CANADIAN THESES ON MICROFICHE

THÈSES CANADIENNES SUR MICROFICHE

National Library of Canada
Collections Development Branch

Canadian Theses on
Microfiche Service

Ottawa, Canada
K1A 0N4

Bibliothèque nationale du Canada
Direction du développement des collections

Service des thèses canadiennes
sur microfiche

NOTICE

The quality of this microfiche is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Previously copyrighted materials (journal articles, published tests, etc.) are not filmed.

Reproduction in full or in part of this film is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30. Please read the authorization forms which accompany this thesis.

AVIS

La qualité de cette microfiche dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

Les documents qui font déjà l'objet d'un droit d'auteur (articles de revue, examens publiés, etc.) ne sont pas microfilmés.

La reproduction, même partielle, de ce microfilm est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30. Veuillez prendre connaissance des formules d'autorisation qui accompagnent cette thèse.

THIS DISSERTATION
HAS BEEN MICROFILMED
EXACTLY AS RECEIVED

LA THÈSE A ÉTÉ
MICROFILMÉE TELLE QUE
NOUS L'AVONS REÇUE

Canada

NL 339 (r. 86/06)

(117)

0-315-01311-6

National Library of Canada
Canadian Theses Division
Ottawa, Canada
K1A 0N4

Bibliothèque nationale du Canada
Division des thèses canadiennes

PERMISSION TO MICROFILM — AUTORISATION DE MICROFILMER

• Please print or type — Écrire en lettres moulées ou dactylographier

Full Name of Author — Nom complet de l'auteur

LEROY M. SUMNER

Date of Birth — Date de naissance

JULY 15. 1950

Country of Birth — Lieu de naissance

ANDROS - BAHAMAS

Permanent Address — Résidence fixe

P.O BOX N-9269
NASSAU N.P.
BAHAMAS

Title of Thesis — Titre de la thèse

AN INVESTIGATION OF THE NEDELSKY METHOD OF STANDARD SETTING IN THE ALBERTA ACHIEVEMENT TESTING PROGRAM

University — Université

UNIVERSITY OF ALBERTA

Degree for which thesis was presented — Grade pour lequel cette thèse fut présentée

M. ED.

Year this degree conferred — Année d'obtention de ce grade

SPRING 1985

Name of Supervisor — Nom du directeur de thèse

DR. T. O. Maguire

Permission is hereby granted to the NATIONAL LIBRARY OF CANADA to microfilm this thesis and to lend or sell copies of the film.

The author reserves other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without the author's written permission.

L'autorisation est, par la présente, accordée à la BIBLIOTHÈQUE NATIONALE DU CANADA de microfilmer cette thèse et de prêter ou de vendre des exemplaires du film.

L'auteur se réserve les autres droits de publication; ni la thèse ni de longs extraits de celle-ci ne doivent être imprimés ou autrement reproduits sans l'autorisation écrite de l'auteur.

Date

November 19th 1984

Signature

LEROY M Sumner

NL-91 (4/77)

THE UNIVERSITY OF ALBERTA

AN INVESTIGATION OF THE NEDELSKY METHOD OF STANDARD SETTING IN THE ALBERTA ACHIEVEMENT TESTING PROGRAM

by

(C) LeRoy M. Sumner

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES AND RESEARCH

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE

OF Master of Education

DEPARTMENT OF EDUCATIONAL PSYCHOLOGY

EDMONTON, ALBERTA

Spring, 1985

THE UNIVERSITY OF ALBERTA

RELEASE FORM

NAME OF AUTHOR LeRoy M. Sumner

TITLE OF THESIS AN INVESTIGATION OF THE NEDELSKY METHOD OF STANDARD SETTING IN THE ALBERTA ACHIEVEMENT TESTING PROGRAM

DEGREE FOR WHICH THESIS WAS PRESENTED Master of Education

YEAR THIS DEGREE GRANTED Spring, 1985

Permission is hereby granted to THE UNIVERSITY OF ALBERTA LIBRARY to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only.

The author reserves other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without the author's written permission.

(SIGNED) .....

PERMANENT ADDRESS:

P. O. Box N-9269

NASSAU N. P.

BAHAMAS

DATED November 19th 1984

THE UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES AND RESEARCH

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research, for acceptance, a thesis entitled AN INVESTIGATION OF THE NEDELSKY METHOD OF STANDARD SETTING IN THE ALBERTA ACHIEVEMENT TESTING PROGRAM submitted by LeRoy Sumner in partial fulfilment of the requirements for the degree of Master of Education in EDUCATIONAL PSYCHOLOGY.

Thomas O. Maguire

Supervisor

Don Massey

V. R. Nybo

Date. NOVEMBER 7th 1984

DEDICATION

This thesis is dedicated to my family, and especially in memory of my dear sister Cardell N. Sumner who died during the writing of this thesis. Though your eyes are now closed in perpetual sleep, the fond memories and helpfulness will always linger in my mind. I will always love you.

ABSTRACT

The use of normative data to help judges in their deliberations has recently become a central issue in the debate on standard setting. The pervailing contention is, such information would enable judges to set more reasonable expectations(passing scores) for students as they would no longer rely entirely on their individual varied experience and knowledge of students and methods.

This study investigated empirically how content specialists working as judges, performed in a judgmental standard setting exercise; and examined the influence of normative data on judges' opinion during this iterative judgmental process. A modification of the Nedelsky (1954) method was used with the Alberta grade 9 Social Studies Achievement test.

The study was an ex-post facto design. A total of 23, 579 student in 564 schools from 104 jurisdictions wrote the test. Seventy-one grade 9 Social Studies teachers selected from the jurisdictions participated in the study as judges (raters). Data collected by the Student Evaluation Branch of Alberta Education was reanalized in an attempt to provide answers to six questions salient to the performance of content specialists in a normative standard setting exercise.

The analysis of the data revealed that judges - content specialist - do not set the same standards. The first session standard or cut-off score was significantly

different from the standard for the second session. It was further revealed that judges do change and some judges changed more than others. This suggest therefore that the normative information provides during the the second secession seem to have had varying influence on judges judgments. Finally, the analysis further revealed that judges interacting with normative data to aid their deliberation of setting borderline cut-off score do tend to use the information in favour of the borderline students.

The results have implications for the use of normative data as supplemental information in a standard setting exercise as well as for further study and caution for those using standard setting methods.

ACKNOWLEDGMENT

I wish to extend my sincere appreciation and indebtedness to all those who assisted in bringing this research to a successful conclusion. First, I would like to express my sincere gratitude to Dr. T.O. Maguire, my M.Ed program and thesis supervisor for his supervision, guidance and patience displayed throughout the course of my study. Sincere appreciation is also extended to Dr. V. Nyberg and Dr. D. Massey, members of my thesis committee, for their advice and suggestions.

I also wish to acknowledge the assistance of Dr. Dwight Harley for his assistance with the computer program and statistical analysis.

A debt of gratitude is also owed to the Personnel in the Student Evaluation Branch of Alberta Education. In particular I wish to thank Dr. L. Symyrozum for giving me access to the data, and to Mr. David Wassaman for the retrieval of data used in this study.

I remain especially grateful to the Canadian Universities and Colleges and the Government of the Commonwealth of the Bahamas whose joint sponsorship made it possible for me to pursue the study.

Lastly and of paramount importance, I am indebted to my wife, Ruth, for her understanding, encouragement, and moral support in those moments of fatique and frustration experienced throughout the course of my study; and my children, Kareem, Michaela, Nikita and Candilaria whose

presence reminded me that "life is not all books".

Table of Contents

Chapter Page

1. INTRODUCTION 1

1.1 Background to the problem 1

1.1.1 Purpose of the Study 5

1.1.2 Importance of the Study 8

1.1.3 Rationale for Using the Nedelsky's Method ..8

1.1.4 Definition of Terms 10

1.1.5 Summary 11

1.1.6 Organization of the thesis 12

2. REVIEW OF THE LITERATURE 13

2.1 Historical Background: Standard Setting Methods .13

2.1.1 The Use of Normative Data in Standard Setting 20

2.2 External criterion 21

2.3 Distribution of examinees scores 21

2.4 Normative data as supplemental information 22

2.4.1 Review of Some Popular Methods 23

2.5 Judgmental Methods 24

2.5.1 (Angoff, 1971) The Angoff Adaptation 25

2.5.2 Modified Angoff 25

2.5.3 (Ebel, 1972) The Ebel Procedure 26

2.5.4 (Nedelsky, 1954) The Nedelsky Procedure ...28

2.5.5 (Jaeger, 1982) The Jaeger Method 29

2.5.6 The Review Panel Method 31

2.6 Empirical Methods 32

2.6.1 Livingston (1975) - Data-Criterion Measure: 32

2.6.2 Regression & Decision Theoretic Approaches(1972) 33

2.7 Combination Methods 33

2.7.1 Berk (1976) 34

2.7.2 Ziekey & Livingston (1977) Borderline Group Method 34

2.7.3 Zieky & Livingston (1977) Contrasting Group Method 35

2.8 Administrative Methods 36

2.9 Comparative Studies 36

2.9.1 Summary 44

3. THE ALBERTA ACHIEVEMENT TESTING PROGRAM 46

3.0.1 Goal of the Alberta Achievement Program ...47

3.0.2 Interpretation of Test Scores 47

3.0.3 Summary 50

4. METHODS AND PROCEDURES 51

4.0.1 Instrument 51

4.0.2 Method and Procedure 52

4.0.3 Data collection 53

4.0.4 Treatment of Data 54

5. RESULTS AND DISCUSSION 57

6. SUMMARY, CONCLUSIONS, IMPLICATIONS AND RECOMMENDATIONS 83

6.0.1 Results 85

6.0.2 Conclusions 86

6.0.3 Implications 87

6.0.4 Recommendations 88

REFERENCES 91

Appendices 97

LIST OF TABLES

Tables	Description	Page
1	A Classification of Methods for Setting Standards.	99
2	Andrew & Hecht Study.....	99
3a	Stem an Leaf display of Judges Standard Scores for Time 1 Rounded to Nearest Whole Number.....	99
4a	Stem and Leaf Display Judges' Recommended Cut Off Scores Time 2 (Rounded to Nearest Score).....	99
4b	Distribution of Judges Standard Scores for Second Rating.....	99
5	Mean Comparisons Between Judges Standard Score for Time 1 and Time 2 Rating.....	99
6b	Back-to-Back Stem and Leaf Display of Judges Standard Score for Time 1 and Time 2 (Rounded to Nearest Score).....	99
8	Summary of Judges Overall Level and Individual Change From Time 1 vs Time 2.....	99
9	General Group Changes.....	99
10	Modal Summary of Judges Item Ratings Difficulty and Shifts for Both Sessions.....	99
11	Distribution of Judges rating of Item Relevance to Degree of Estimated Change.....	99
12	Histogram of Contingency of Coefficient of the Relationship of Items Relevance to Judges Estimated Difficuly.....	99
13	Histogram of Contingency of Coefficient of the Relationship of Estimated Difficulty.....	99
14	Chi Square Relationship Between Judges Estimated Rating of Item Difficulty to Item Relevance.....	99
15	Relationship of Judges First Ratings of Item Difficulty to Observed Test Item Difficulty.....	99
16	Frequency Distribution of Relationship of Judges First Estimated Ratings of Item Difficulty to	

Corresponding Observed Test Item Difficulty.....99

17 Chi Square Relationship of Judges Estimated Item Difficulty Ratings to Observed Test Item Difficulty for First Rating.....99

18 Relationship of Judges Second Ratings of Item Difficulty to Observed Items.....99

19 Frequency Distribution of Relationship of Judges Second Estimated Ratings of Items to Observed Test Item Difficulty.....99

1. INTRODUCTION

1.1 Background to the problem

Since the concept of criterion-referenced testing was introduced during the 1960's (Glaser, 1963, Popham & Husek, 1969) much attention and debate have focused on criterion-referenced measures and setting cutting scores for such measures.

Several alternative procedures for setting standards have received attention in the literature. Extensive descriptions of the properties of the methods, their underlying assumptions, the purpose(s) each addresses and the general procedure to follow when setting standards are many (Poggio et al., 1981). Teachers having very little experience as well as the more seasoned and experienced teachers are often required to set reliable and meaningful standards for granting special honours for excellence and determining pass/fail in a testing program. Until now, most work in standard setting has been a priori. That is, judgements are made without knowing how the student actually performed. Many researchers now feel that the current practice of setting standards without the use of some normative data is like shooting in the dark at a black object. Shepard (1976) and Jaeger (1976, 1982) protest that judges ought to be provided with normative data in their deliberations. While the norms are not automatically the standards, with the aid of normative information, judges can

be expected to establish more reasonable expectations.

The iterative procedure requires judges to rate items on a number of occassions. After each session they are given summary information and provided with a cummulative distribution of the recommended standard. They can avail themselves of the opportunity to reconsider their initial and subsequent item-by-item recommendations, and to make changes if they so desire. At the conclusion of the final rating session, the results for each judge are pooled and a median is computed for each judge. The minimum median across all judges is selected as the standard.

Jaeger (1982) conducted an iterative structured judgment process for setting standards. In his study on the North Carolina High School Competency test in Reading and Mathematics, he asked judges to rate whether every regular high school graduate should be able to answer each of the items. Scores were determined by adding the number of affirmative judgments, then the results was given to each judge together with the distribution of assessments provided by the other judges. Judges were then allowed to adjust their assessments. Later in the process, student performance data was supplied and the judges were again allowed to change their assessments. The results of the Jaeger study indicated that the student performance data has an important role to play in structured judgment procedure.

In the most recent study to-date, Cross et al (1984) compared the effect of normative feedback on three methods.

The Nedelsky, Angoff and Jaeger methods were compared in the National Teacher Examinations(NTE),in order to establish a minimum,yet valid standard for teachers certification in the state of Virginia. The scope of this study was examined under three levels of normative information provided to judges in a three consecutive review sessions. The results corroborate the findings of the Jaeger(1982) study that normative feedback is useful to judges in a standard setting exercise. The results further indicate that while the Nedelsky's method produced the most lenient standard, the Angoff's standard was more defensible and the normative feedback had its effect with it. However, while the supplemental data appeared to influence the decisions, we still do not know the specific effects that it has. Neither do we know when, i.e., at what point in the process, this data should be given to judges engaged in the exercise.

The Student Evaluation Branch of Alberta Education conducted its initial Achievement test in June of 1982. After the results were released, it became clear that the reported results were more descriptive than judgmental which posed some problems for Provincial and jurisdictional interpretation. Furthermore, as Maguire (1982) pointed out, results as reported did not provide any standards against which Provincial or jurisdictional performances could have been measured.

The following year, 1983, the department again administered another round of achievement testing. This

time, it engaged in some judgmental process of standard setting for the various subjects. The Nedelsky (1954) procedure was used for the multiple-choice section of the grade 9 Social Studies examinations. It has been suggested by several authors, (Shepard 1976, 1980; Jaeger 1976, 1982; Rowley 1982) that the task of setting standards would be unreasonable without the use of some normative test statistics. They further recommend that the task not only use normative test statistics but the process also be an iterative one, in which teachers interact with the normative data and make modifications to their item-by-item recommendations if they thought it necessary and desirable. Normative as used in this context means the supplemental use of performance information in the standard setting exercise and not use of the performance data as an external criterion or standard. Additionally, it is the view that since teachers are involved with the subject content and students, they should be knowledgeable about the expected level of achievement so they should be the ones engaged in a standard setting exercise.

Although the Nedelsky's procedure has been around for some time, there has been no attempt to modify it to form an iterative standard setting process using normative test data. Hence the need to investigate and analyse the judges' ratings and the influence of test data in an iterative process on teachers' judgment.

1.1.1 Purpose of the Study

Determining standards for tests Jaeger said, is "both vexing and difficult". He went on to assert that it is perhaps "the weakest link in some otherwise defensible programs" (p. 461). An existing problem in the Alberta Achievement testing program is the determination of standards against which to interpret test results. The Alberta context is not a minimum competency testing context, and so methods for setting standards have to be useful not only for determining pass-fail, but also for granting special honours for excellence. It follows then that if teachers are knowledgeable in their content area, as well as with the abilities of their students, the standard arrived at should therefore be less capricious and more reliable than if they are not. This is extremely important since children from the same jurisdiction will be affected by different teachers.

Several authors (Jaéger, 1982; and Shepard, 1976, for example), recommend iterative procedures. Judgments are fundamentally human and therefore, they are subjective. The purpose of the standard setting exercise is to provide targets for performance and reinforcement for success. The reinforcement value is at least partly related to the level of the standard, so that standards must be high enough to be reinforcing, but low enough to be attainable. Finally, standards are socially, politically and economically influenced. If standards are too high and too many students

fail, then there will surely be a public outcry about the quality of the schools and the unreasonableness of the standards. Furthermore, if a state or province is committed to remediation the cost of remediation could be very high. By the same token, if standards are set too low, then the program becomes meaningless, and if people become aware of the ridiculously low standards, they will again present an outcry about the quality of the schools. Yet they must seem appropriate and defensible to the various publics: parents, employers, students, current teachers and subsequent teachers.

Since judgments must satisfy all of these purposes, and since human assessment is fallible, it would seem appropriate to evaluate procedures that allow standards to be modified by test results. The Nedelsky procedure allows for this.

The Nedelsky (1954) method is used exclusively with multiple-choice tests and focuses on an analysis of the wrong-answer options in such tests. Using this procedure, a panel of judges is given copies of the examination. Each judge is asked to review independently every item and cross out those alternatives that he or she considers a borderline student should be able to eliminate as incorrect. The minimal passing level (MPL) for that item is the reciprocal of the number of remaining alternative(s). The standard is then estimated by averaging each judge's item passing level. These values are then averaged over judges to get the final

standard.

A search of the literature revealed that no study has been done using a modified procedure such as the Nedelsky method incorporating the use of normative data and an iterative process styled after the Jaeger format to set standards. The Alberta Achievement testing program was an ideal venue for such an investigation.

The purpose of the study then was to investigate and evaluate modifications of the Nedelsky method of standard setting with the aid of normative data using an iterative process. More specifically the following questions were investigated:

1. Do judges using the same test content data and the same method set the same or different standard?
2. What effect does test analysis data, when used in an iterative standard setting context, have on influencing judges' opinion ?
3. Do some judges change more than others after receiving the normative feedback ?
4. If judges classify items as essential, are they more or less resistant to change than if they classify items as important, acceptable or questionable?
5. How do the judges ratings of EIAQ classification relate to the test item statistics?
6. If judges do change, do they tend to change towards congruence? That is, do they change toward making the item more like the observed difficulty or less like the observed

difficulty?

1.1.2 Importance of the Study

The Alberta Achievement Testing Program results were released in November of 1982. "In some cases, implicit interpretations of the results were made, but for the most part the reports were characterized by description rather than judgment. It did not provide any standards against which provincial or jurisdictional performances could be measured (Maguire, 1982, p.3)." The study is therefore important for several reasons:

1. It will give some indication of the validity and stability of teachers' judgments when given actual test statistics as supplemental data to aid in setting standards using a judgemental method.
2. It is hoped that the result will contribute to the body of literature on standard setting methods, especially as it relates to the use of supplemental data during an iterative process.
3. The critique and integration of the various methods will hopefully aid practitioners in selecting a standard setting method.

1.1.3 Rationale for Using the Nedelsky's Method

Standards should be set low enough to be attainable but yet be , at the same time, high enough to be defensible. Shepard (1976) makes the point: If standards are to be used

to make decisions about individuals, then one set of criteria is needed. The criterion should either be the most stringent or the most lenient of those proposed depending on which type of error is more serious in that situation. If false negatives would be more costly to individuals and society, then the standards should be lower than in the instances when false positives must be screened out (p.31). Popham on the other hand said that it is better to set low standards and adjust upwards rather than set high standards and adjust downwards. The Nedelsky method speaks well to this rationale. For example, in all studies that compare Nedelsky and other procedures, the Nedelsky procedure produces a lower standard. But, the formula allows for upward adjustment if necessary. The Nedelsky focuses on items and alternatives, (i.e. parts of items), and may, to some extent, model the behavior of a student as she or he attempts a difficult item. According to Shepard (1980), the Nedelsky procedure should not be used unless elimination of wrong answers is clearly consistent with how a minimally competent person would answer the test. She goes on to say that "In most minimum competency testing situations items have been written so that the hardest discrimination (choosing between the right answer and the next best answer) reflects the minimum competency to be addressed." It seems likely that if items are developed so that the alternatives represent varying degrees of knowing then the Nedelsky procedure, or some refinement of it would be the procedure

of choice. This is particularly true when we are dealing with a test that measures a single facet of achievement. Furthermore the Nedelsky formula is such that once a cut-off s core has been determined, one can then adjust it up or down depending on the number of students one wishes to fail or pass.

1.1.4 Definition of Terms

Acceptable items:

Acceptable items in this context refers to items that are rated as being not directly applicable to problem solving. Successful performance on these items is neither essential nor impor tant to the goals of the course.

Cut off Score

A cut-off score is that score which differentiates individuals who "pass" or "fail" a test. In the literature it is used interchangably with standard.

Essential Item

For the purpose of this study, an essential item is defined as testing content in which a candidate is expected to have a high degree of knowledge. The item refers to a fundamental piece of information or principle the student ought to know. Failure to know or get this item correct would indicate a serious gap in the student's knowledge of that subject.

Important Item

Important item refers to items that are important for the student to know but do not have the same priority as items

classified as essential. A student not getting this item correct will be considered deficient in that content.

Judges

Judges are defined as teachers used in the standard setting method to rate the items in a test and make a determination (judgement), of which item(s) or distractor(s) a barely passing student would recognize as incorrect, thus ultimately deciding the minimum standard or cut-off score acceptable to pass, graduate or award certification for the concerned audience.

Standard

A standard is a point on a test score scale that is used to 'sort' examinees into two categories that reflect different levels of proficiency relative to a particular objective measured by a test. It is the lowest score which a student may obtain and is acceptable as a 'pass' for promotion or certification.

Standard Setting

The process by which the standard or cut-off score for a minimal competency or objective test is established. Standard setting is essentially a judgmental processs.

1.1.5 Summary

This chapter provided a basic overview of the frame of reference for the study. It pointed up the vacillating state of affairs with respect to setting cut-off scores or standards against which institutions might use as a measure

to interpret students' performance.

Many methods for setting standards were advanced and tried. These were subsequently criticized on the ground that they were too judgmental and arbitrary. Researchers and educators felt it was not fair to expect judges to set standards in the absence of normative test data. Thus, it was proposed that normative information should be available to judges when engaged in setting standards. The whole debate then has come to rest (at least for the moment) on using a combination of normative data along with a method to derive a reliable and less capricious standard or cut-off passing score.

1.1.6 Organization of the thesis

The remaining chapters of this thesis are organized as follows: Chapter II is a review of some of the more popular standard setting methods as well as a review of some of the comparative studies. Additionally it touches on the Alberta Achievement testing program and the reasons for engaging in standard setting. The third chapter looks at the methodology, procedures, and instrumentation utilized in the study, as well as the treatment of the data. Chapter four looks at the findings and analysis of the data and questions proposed in the study. The fifth chapter discusses these findings question by question. The final chapter provides a summary, conclusions and recommendations resulting from the investigation.

2. REVIEW OF THE LITERATURE

This chapter briefly traces the issue of standard setting, the move away from norm referenced test interpretation to a criterion referenced mode of interpretation, and the debate surrounding this change. Additionally the chapter reviews the more popularly utilized standard setting procedures while critiquing some of the comparative studies associated with these methods. Finally, it looks at the standard setting issue as it relates to education in the the Province of Alberta and its jurisdictions.

2.1 Historical Background: Standard Setting Methods

The new surge and shift away from traditional (norm-referenced) testing is quite recent. About the turn of the 1960's, norm-referenced test interpretations of the grading system were seen as too competitive and therefore inevitably produced winners and losers. It was further seen as less than ideal for providing the desired kind of test score information that was being requested. The feeling was widespread that students should be evaluated against their own standards, or against some 'absolutes' and without comparison with other students as was the practice of the traditional norm-referenced method.

The public too was demanding evidence of school program effectiveness as well as student competency performance in basic subject areas; especially for high school graduation

and promotions. Criterion Referenced Testing (CRT) was seen as the alternative. The concept was introduced and made popular around the 1960's by such noted scholars as Glaser and Popham (Poggio et al., 1981).

Since then acceptance and application of the concept has shown tremendous growth. This 'new' alternative has been hailed as a kind of 'shot in the arm' for the ailing student performance in subject areas. It was supposed to provide the kind of test score information not only necessary for making a variety of individual and programmatic decisions arising from the objective based instructions, but more importantly, it was to bring to a halt what many people complained about; a devaluation of the High School diploma (Popham, 1981).

The call for criterion-referenced testing was very dramatic to say the least. Berk (1981) asserts that it was perhaps the most dramatic change among the many during the testing era of the 1960's. The change was dramatic for several reasons incuding the fact that it called for a new approach to the writing of objective based items that would validly measure achievement in specific domains. More important, and the area that caused much concern, was the issue of defining and setting performance standards or cut-off-scores for such criterion reference testing; even though the notion of cutting scores was not originally part of the criterion referenced movement. In spite of this, the notion of setting cutting scores has permeated all levels of educational testing and indeed has become the central issue

in standard setting.

Maguire (1983) has done an extensive critique on the standard setting issue. He noted that prior to the 1960's, the issue of standard setting a to be directly answered in one of two ways, either by defining a proportion of people who would pass a test, or by specifying a passing score (commonly expressed as a percent). While both procedures were arbitrary, neither was haphazard. Anyone with considerable experience in test design is capable of producing questions whose difficulty levels do not vary much from one occasion to another. He noted further that with large populations the average ability level remains fairly steady over time so that whichever procedure was used, the standards were probably the same from one year to the next.

Maguire surmised further that during the 60's and 70's, two factors seemed to have influenced standards. The first he noted was the general direction toward local autonomy in education which meant the classroom teachers were required to set standards at the classroom level without the benefit of knowledge of the performance of the wider population of students. Secondly, the influence was the philosophical pressure from the objectives-based testing movement. Consequently, "advocates urged teachers to abandon the practice of setting standards by comparisons with other students, and adopt criterion referencing(Maguire, 1982, p.5)". Properly written educational objectives were to include the standard or level of performance required for a

pass.

The analysis is over simplified, yet it is true that the idea of standards which had previously been accepted became a focus for debate. The issue was so important a special issue of the Journal of Educational Measurement (Winter, 1978), was devoted entirely to standard setting. Maguire critiqued three articles from that issue he thought deserve special review because they focus the debate. The first article was by Gene Glass.

Maguire observed that in a well researched article, Glass (1978) traced the evolution of the notion of "criterion" as "standard", pointing out that criterion referencing need not have involved the specification of a standard performance. Behavioral objectives that state a standard such as "..... the student must correctly solve 7 equations within thirty 'minutes", were described as pseudo quantification. More importantly, they were a meaningless attempt to make objective a matter that is essentially judgmental and, in most cases, confused with what is essentially a performance comparison of norm referenced standard.

Glass identified and criticized six classes of techniques for determining criterion scores.

1. Performance of others. The passing score is set as the score achieved by a known population. For example, the Medical Council of Canada sets a standard for non-Canadians who take the licensing exams as a particular percentile of

the distribution of scores made by Canadian graduates. This is clearly a norm referenced standard.

2. Counting backwards from 100%. If a particular skill is intended to be basic, then the desired performance level is 100%. However, allowance is usually made for clerical errors, inattention, etc., and the level is moved down to 95% or 80% or some other value that appears to have no particular justification. The 80% level seems very common among criterion referenced test advocates but no rationale appears to have been enunciated.

3. Bootstrapping on other criterion scores. While this technique has seldom been used, it is often suggested that the passing score be established to fit the criterion of an already established competent person in the area. This is virtually identical to the problem of predictive validity in classical test theory, and merely pushes the problem of definition of competence from the original test, back to a second test.

4. Judging minimal competence. Glass discusses the various judgmental procedures for establishing standards. His review of the literature is incomplete, but the critique of the notion of "minimal competence" is compelling in that it is difficult if not impossible to define "minimal" in any unequivocal or psychologically meaningful sense. Nonetheless, the techniques described hold some promise for the purpose of the Student Evaluation Branch.

5. Decision theoretic approaches. A great deal of time has

been spent in viewing the cutting score problem as a problem in decision theory. All procedures founder on the need for having an external criterion which has a meaningful competence level. Essentially this strategy, bootstrapping and operations research, are special cases of the treatment of validity in classical test theory. They eventually come down to a norm referenced base.

6. Operations research. An attempt is made to find a cutting score that maximizes some valued commodity. Here the commodity might be future achievement for example. Generally the outcomes are weighted composites since any skill leads to several others. The weights are arbitrary, and since the cutting scores will be largely dependent on the criterion weights chosen, the method appears at best, to obscure the essential judgmental nature of the criterion.

Finally, Maguire in his paper noted that Glass, makes the points that most standard setting is arbitrary, or comparative. There is no self-evident standard. In Glass' opinion the most attractive alternative is to observe rates of performance and see whether they go up or down. Of course according to Maguire (1982) "what he does not discuss is the need for boundaries that tell the observer when a downshift should become a matter of concern and when it is tolerable".

The second article critiqued by Maguire was that of Scriven. He noted that Scriven(1978) in the same issue of the Journal of Educational Measurement, comments on the Glass paper and indicated that it is unreasonable to talk

about mastery/nonmastery as a black and white distinction. At the border there are areas of grey, and that the assignment rules are arbitrary as you move away from it. Additionally, he criticizes Glass for implying that standard setting is an all or nothing affair. Anyone in an applied science field ought to automatically investigate the consequences of standards in order to modify them for subsequent use. In other words, we should learn from our experience. To quote:

> To put the matter bluntly, the answer to Glass does not lie in present practice (nor does it lie in surrender) but rather in elements missing from the whole picture including better procedures for calibrating and training judges, for synthesizing subtest scores, and - especially - in needs assessment (p. 274).

Maguire contend that the third useful article from the issue on standards is Hambleton (1978). Hambleton distills Glass' criticisms to two statements: (1) all methods of determining cut-off scores are arbitrary but it would be safer if less arbitrary methods were used; and (2) the use of examinee gain scores is preferable because they are less arbitrary.

In Glass' article, criticism of standard setting procedures was based in part on a particular piece of research in which two standard setting methods were conducted by Andrew and Hecht (1976). This article (to be

discussed in more detail later) described an experiment in which judges were asked to set cutting scores using two separate procedures. The results showed large differences in the cutting points set. Hambleton points out that the results are not surprising since the approaches used were conceptually different, but that different groups of judges using the same approach were remarkably consistent.

Responding to the first criticism, Hambleton suggests that the research offers some hope for developing methods that are consistent and useful, albeit arbitrary to some extent. With regard to the second criticism, at the level of the individual, gain scores are remarkably unreliable, and their use at that level may be harmful.

While Glass' article makes several valid points, Maguire felt the rebuttals by Scriven and Hambleton seem sufficiently compelling to warrant further research into standard setting.

2.1.1 The Use of Normative Data in Standard Setting

Arguments for the use of normative data as an aid (reality check) to standard setting have been advanced by such researchers as Shepard (1976), Jaeger (1976) and Conaway (1976, 1977). It has been conveyed that the normative data could be put to one of three uses: As external criterion, as decision based on distribution of examinees' scores and as supplemental information.

2.2 External criterion

The contention is that use could be made of the normative performance of some external "criterion" group. Jaeger (1976) provides as an example, the Adult Performance Level (APL) test by Palm Beach County, Florida Schools. Test performance of groups of "successful" adults were used to set standards for high school students. This procedure, however, was criticized on several counts. Jaeger (1978) himself, correctly pointed out that one of the dangers is the changing nature of society and that standards change with societal changes. Therefore, standards set on adult performance may not be relevant to high school students. Burton (1978) points out further, that relationships between skills in school subjects and later success in life are not readily determinable, hence observing the degree of achievement on the test of some "successful" norm goup makes little sense. Jaeger (1978) extended this view when he noted that there are no empirically tenable survival standards on school based skills that can be justified through external means. In other words, using an external criterion is not a recommended use of normative data in the standard setting exercise since it has little or no utility.

2.3 Distribution of examinees scores

A second way of proceeding with normative data is to make a decision about a standard based exclusively on the distribution of scores of examinees who take the test. Such

a procedure, it is claimed, circumvents the "minimum test score for success in life problems", but the procedure is still not useful for setting standards. Glass (1978) argued against putting normative data to this use because this is very similar to the traditional norm referenced approach. In other words, in this case, an individual passing or failing a test depends upon the other individuals taking the test.

2.4 Normative data as supplemental information

The use of normative data as supplemental information to aid standard setting has found more support than the first two, which according to the critics, represent nothing more than poor forms of norm referencing. However, the use of Normative data as supplemental information in the standard setting process has found much support even among the critics. The consensus is that, as part of their deliberations, judges should have normative data at their disposal, as a kind of reality check. With its use as supplemental aid they could be expected to establish more defensible cut- off scores, since they would know how a group of real test-takers performed on the test. Livingston and Zieky(1982) suggested "..if judges idea of a borderline test-taker is someone whose performance approaches or exceeds that of the average actual test-taker, their standards may be unrealistically high(p.57)." Jaeger (1976, 1982), Shepard (1976) have further suggested that not only should the process of setting standards use normative data

but recommended the procedure to be an iterative process, a process that allows judges to change or not change their opinion after being exposed to additional information. Livingston and Zikey(1982) however perscribed that you do not share the normative information with the judges until after they have made their initial judgments. This point will be reexamined again, but first it is useful to review the research involving the various procedures.

2.4.1 Review of Some Popular Methods

Although there is a plethora of procedures for setting standards, the general consensus is that these methods can be organized under three main classifications: judgmental, empirical, and combination (Jaeger, 1976, Stoker, 1976, and Hambleton, 1980). Berk (1980) gives a very good classification scheme of these methods for setting standards. A description of the more promising and common methods will be discussed.

Table 1: A Classification of Methods for Setting Standards

Judgmental Methods	Empirical Methods	Combination Methods
Item Content	**Data-Criterion Measure**	**Judgmental-Empirical**
Angoff(1971)	Livingston (1975)	Borderline groups (Zieky & Livingston, 1977)
Modified Angoff (ETS, 1976)	Livingston(1976)	Contrasting groups (Zieky & Livingston, 1977)
Ebel(1979)	Regression & Decision Theoretic Approaches Kriewall(1972)	
Nedelsky(1954)		Criterion group (Berk, 1976)
Jaeger(1978 & 1982)		
Review Panel Method (B.C.)		

2.5 Judgmental Methods

The judgmental methods of setting standards are also referred to as 'proximal' or 'direct' models. These methods deal with item content in one way or another. Researchers argue that all judgmental methods require data to be collected from judges for setting standards or making judgment as to who is minimally competent and should therefore pass or fail. Jaeger (1976) contends, "all that varies is the proximity of the judgment determining data to the original performance". In these methods, items are inspected, with the level of concern being how minimally competent individuals would perform on each test item. A

description of the more promising judgmental (proximal) models as classified Berk (1980) is shown below.

2.5.1 (Angoff, 1971) The Angoff Adaptation

This method stresses item-by-item judgments. This approach requires that judges come up with estimates of how difficult the individual items are. Each judge examines each item and states the probability that a "minimally acceptable person" would answer each item correctly. Alternatively, the judges could think of a number of minimally acceptable people and estimate the proportion who would get the item correct. A probability of 1.00 is given to that item. If, by the same token, the item is judged to be so difficult that it would be unlikely that a borderline student could answer it correctly, then raters might assign it a low probability. This low probability, however, should never be lower than would result from a raw chance based on the number of alternatives in the item. The sum of the proportions would constitute one judge's estimate of a cutting score. Estimates are pooled across judges to reach a final result.

There are various refinements to this procedure.

2.5.2 Modified Angoff

Educational Testing Service (ETS) 1976, concocted a modification of the Angoff method. It rationalized the necessity for modification on the grounds that the task of assigning probabilities may be overly difficult for the

items to be assessed. ETS (1976) subsequently utilized a seven-point scale on which certain percentages were fixed. Judges were asked to estimate the percentages of minimally knowledgeable examinees who knew the answer to each test item. The procedure included the use of a scale for judges' estimates (5, 20, 40, 60, 75, 90, 95, DNK) and using different summary procedures to combine judgements (mean, median, trimmed mean, etc. a standard was derived. The Insurance Licensing Examinations is often cited as an example. Sixty was used as the centre point on this exam, and the other optio were then spaced on either side of 60.

2.5.3 (Ebel, 1972) The Ebel Procedure

Judges perform a series of tasks in the Ebel Procedure. To begin, they classify each item into a two dimensional table of difficulty and relevance. The categories of relevance are: essential, important, acceptable, and questionable (AEIQ). The three difficulty levels are: easy, medium, and hard. After assigning each item to a cell, the judges are asked to assign a percentage to each cell indicating the proportion of items in that cell that a minimally competent candidate should be able to answer. The number of items in the cell is then multiplied by the proportions, and the results summed over cells to reach a standard.

The various modifications of the Ebel method are many. The number of categories used may vary from application to

application. Sometimes pilot data are used to classify the items according to difficulty. Alternatives to the relevance continuum (e.g. taxonomic level) have been used.

It has been documented by the various researchers (Glass 1976; Burton 1976; Shepard 1976) that all of the methods appear to have numerous potential drawbacks and thus threaten the validity of the findings. In the Ebel (1972) method, for example, potential users should be knowledgeable about some of these drawbacks before selecting it as the model of choice.

Hambleton et al. (1978b) pointed out that Ebel offers no prescription for the number or type of descriptions to be used along the two dimensions--difficulty and relevance. This is left to the discretion of the individuals rating the items. It is therefore conceivable that a different set of descriptions applied to the same test would give a different standard.

Secondly, the process is predicated upon the decision of judges and while the standard could be called absolute in that it is not referenced to a score distribution, it cannot be called an objective standard.

A third caution came from Measkauskas (1976). He warned that judges have to simulate the decision process of the examinee to obtain an accurate judgment and thus set an appropriate standard. The difficulty with this is that the judges might be apt to systematically over-simplify the examinee's task. The reasoning being, that judges are

supposedly more knowledgeable than the minimally qualified examinee, and thus they are not forced to make a decision about each of the alternatives. It therefore allows the judges (unfortunately) to ignore some of the finer discriminations that an examinee needs to make and would consequently result in a standard that is difficult to reach.

2.5.4 (Nedelsky, 1954) The Nedelsky Procedure

Judges examine each alternative in each item and indicate which alternatives the minimally competent student should reject as being incorrect. The minimal passing level for that item is the reciprocal of the number of alternatives left. The standard is estimated by averaging each judge's item passing level. These values are then averaged over judges to get the final standard P. Recognizing that there would be a grey area around the standard, Nedelsky suggested that the value of P be adjusted to

$$P(1) = P + ks.$$

The value s is the standard deviation of judges' standards, and k is a constant which determines how many of the borderline candidates pass or fail. Assuming that borderline candidates have scores that are normally distributed with a mean of P and a standard deviation of s, setting k to 0 would pass one half of the borderline group, setting k to -1 would pass 84% of the borderline group.

There have been few modifications of the Nedelsky Procedure. In most applications the ks term is ignored (equivalent to k=0). As we shall see in the research to be reviewed, the context in which the Nedelsky Procedure is used has been varied, so that the mind set which the judges bring to the task may differ from one application to another.

2.5.5 (Jaeger, 1982) The Jaeger Method

Jaeger describes his procedure as a structured, multiphase Delphi procedure. In the example he provides, three groups of judges were used, registered voters, high school teachers and principals and high school counsellors. The process was as follows:

1. Each group was given an overview of the problem.

2. Each judge completed the test, but was not given the results until recommendations had been made (after step 8).

3. Each judge was asked to indicate the necessity of each item, by answering the following question: "Should every regular high school graduate be able to answer this item correctly? (Yes or no.)

4. The number of "yes" items was computed for each judge, and the total constituted the standard for that judge.

5. A distribution of standards was produced and given to the judges and an explanation provided to assist them in

comparing their own standards to the distribution. They were also given a table containing the percentage of students who had answered the item correctly in the previous administration They were encouraged to examine the two handouts in light of the recommendations given at Step 3.

6. The judges were then asked to re-rate each item (Yes, No), keeping the information that they had been given.

7. A new distribution of standards was prepared from the recommendations of the judges.

8. A third rating session was held after the judges had been given the distributions from step 7, and the distribution of test scores obtained by students in the previous administration. They were shown how to interpret standards in terms of the number of students who would pass if standards were set at particular levels.

9. A final distribution of recommended standards was calculated.

An evaluation of the procedure showed little evidence of convergence from one step to the next. While the average standard shifted, the judge variation remained fairly constant. Different groups of judges varied in their estimates, with the voters recommending higher passing scores than others on reading. In mathematics, teachers and voters had almost identical distributions, and both were higher than principals and counsellors.

2.5.6 The Review Panel Method

This method was used in many of the Alberta Minister's Advisory Committee on Student Achievement (MACOSA) tests, and is used in the B.C. assessment program. The procedure used by the B.C. Science Assessment is described here. Interpretation panels consisting of teachers actively involved at the grade level of interest were struck. None of the teachers had been involved in the earlier activities of the assessment (except that their students may have been in the assessment sample). The panels were given the table of specifications and a copy of the achievement items classified according to the learning objectives.

Each panel member began by completing the item and setting a percentage figure for "acceptable" and "desirable" levels of performance for the province as a whole based on the percentage of students who they felt should be able to correctly answer each item. After this, they were given the results for each item in terms of the proportion of students who answered the item correctly and then asked to rate the performance by comparing the results with their previously estimated acceptable and desirable levels. Ratings were made on a five-point scale from weak to strong. An attempt was made to reach a consensus rating through discussion. At a final session panels were asked to develop ratings for each domain (a set of several items testing one area) and to provide interpretive comments and recommendations in light of provincial performances. Again, a five-point scale was

used.

2.6 Empirical Methods

The empirical methods are also known as 'distal' or 'derived' models. These methods involve the use of behavior in a related domain of tasks or some external measure of performance to set the standard of performance for its domain of interest. These methods according to Berk (1980) require the collection of examinee test data to aid in the standard-setting process. The belief is that, by far the largest number of standard-setting methods fall into this category. More of the technical psychometric methods are examined, thus giving the impression that standards can be derived with scientific precision. The argument is advanced by Shepard (1980) that these approaches do not determine a standard, "rather they presume that a standard already exists on an external criterion and merely translate this into a cut-off score on the test "(p. 459).

2.6.1 Livingston (1975) - Data-Criterion Measure:

Livingston argues that this procedure can be used with any suitable criterion measure, not just instructed vs uninstructed as in some methods. He feels that the benefit and cost of a decision are linearly related to the cut-off scores in question when viewing the effect and accuracy of such decision-making. He contends that the cost of a bad decision is proportional to the size of the error made and

the benefit of a good decision is proportional to the size of the error avoided.

2.6.2 Regression & Decision Theoretic Approaches(1972)

There are several empirical procedures using regression and decision-theory approaches. These involve the use of an external criterion and for the reasons described by Glass, seem inappropriate.

2.7 Combination Methods

Several other procedures have been proposed (and often used) for setting standards. Many of them require the collection of achievement data from specific groups of students, and others are entirely ad hoc. Given the requirements of the achievement testing and comprehensive examination programs, they are unlikely to be directly applicable as they stand. Nevertheless, they may play a useful part in the evaluation of other methods and are listed below for that reason.

The combination methods use both judgmental and empirical data in the standard setting process. However, while judgments are required, the judgments are about students, not items, as with such methods as Nedelsky, Angoff, Ebel, etc.. The view held by users of these methods is that judging individuals is likely to be a more familiar task than judging items. In these methods, teachers are the logical choices as judges, and for them the assessment of

individuals is common place.

2.7.1 Berk (1976)

Berk proposed a procedure called criterion groups. This procedure is quite similar if not identical to the contrasting groups method, to be described later. He, however, in attempting to avoid the problem of defining (and judging) mastery and non mastery of the two groups, used instructed and uninstructed groups. Some subjectivity does nevertheless seep into this approach, because in order to classify groups as instructed, the judge must know whether the examinee had received "effective" instruction before he/she can be presumed that they were instructed.

2.7.2 Ziekey & Livingston (1977) Borderline Group Method

This method was first proposed by Zieky and Livingston (1977). It is based on the judgmental identification of students who are considered at or near the borderline. Teachers are asked to select a group of students who they consider to be at the borderline. The test is administered to this group and the median score is taken to be the standard. Popham (1982) provides a procedure to follow for the implementation of this approach:

1. Identify judges who know the student population involved.
2. Have judges discuss what constitutes minimally acceptable performance.
3. Have judges identify borderline students.

4. Administer test.

5. Compute median performance.

User's of this method concede, however, that this approach is plagued by two fundamental problems: The identification of judges being a key factor and being able to clearly identify an adequate sample of only borderline examinees as the other factor. One clear advantage of this method nevertheless, is its simplicity.

2.7.3 Zieky & Livingston (1977) Contrasting Group Method

This is another method proposed by Zieky and Livingston (1977). This is in a way, the converse of the borderline Group method. In the Contrasting Groups method, teachers are asked to classify students into pass and fail or isolate them as absolute masters or nonmasters groups. The test is administered and the score distributions plotted. The point at which the score distribution for the fail or nonmasters group crosses the distribution for the pass or master group is taken to be the standard.

The guideline for the execution of this method is basically the same as for the Borderline Method, with only two other steps:

1. Plot the performance curve for both groups.

2. Set the performance standards based on the intersection of the two curves.

The claim to fame of this approach is that it seeks to minimize the overall classification errors, as the authors

allow for the cut-off score to be moved up or down to reduce , selectively false positive or false negative errors.

2.8 Administrative Methods

Setting standards by administrative method simply means that the cut-off decision is determined by one or more persons holding position of authority. These methods are very commonly employed. The administrative methods cannot be classified as either judgemental or statistical because they have very little structure. Furthermore, the use of such methods gives no reason to believe that they will lead to cut-off scores with appropriate psychometric characteristics nor less arbitrary decision, but rather results in "arm chair" capricious standards that are difficult to defend.

2.9 Comparative Studies

There are not many studies which compare the outcomes of various methods. A brief survey of the literature turned up few comparative studies. Andrew and Hecht (1976) used the Nedelsky and Ebel procedures and individual or consensus judgments on two instruments of 90 items each made up from the odd-even split of a 180-item test used to certify professional workers. The results shown below are the average standards (%) set by the groups under the various conditions.

There was no significant difference between average individual ratings and group consensus. There were no

Table 2

Andrew & Hecht Study

	Nedelsky		Ebel	
GROUP:	A(n=4)	B(n=4)	A	B
Average Individual Standard	50.3	53.7	68.8	68.0
Group Consensus	46.3	51.3	68.4	67.6

differences between groups within a single procedure, but the Ebel procedure produced significantly higher ratings than the Nedelsky procedure.

Skakun and Kling (1980) compared the Nedelsky with two modifications of the Ebel procedure on an examination in medicine. In the first modified Ebel procedure, difficulty levels for each item were given (greater than .80, .30, to .80, less than .30) as were the taxonomic levels of the items (Factual, Comprehension, Problem Solving). In the second modified Ebel procedure, taxonomic level was cross classified with relevance (Essential: - failure represents a serious gap, Important - candidates should know it, Acceptable: - failure does not represent a serious gap). All items had been previously categorized and the judges were required to indicate the proportion of questions in each cell that a barely qualified candidate should answer correctly. When compared to the norm referenced standard already in use, the two Ebel procedures produced similar values. The Nedelsky value was much lower. In addition, the Nedelsky standards had higher variance over judges than the Ebel standards, and the reliability (ANOVA) for the Nedelsky rating was lower.

Rock, Davis and Werts (1980) compared Angoff and Nedelsky procedures for estimating minimal competency and average competency using four groups of judges. The Angoff procedure produced higher values than the Nedelsky procedure, there was more consistency between groups and the group took less time with the Angoff procedure. In addition, the Angoff standards for average competency was closer to the observed mean than the Nedelsky.

Saunders, Ryan and Huynh (1981) compared two approaches, both of which are modifications of the Nedelsky method. They had 180 students rate the questions on a statistics examination that they had taken previously. Procedure 1 required them to place item alternatives into categories such that students at a minimal B level could eliminate them. The score for a minimal B, was calculated as the sum of the reciprocal of the number of alternatives so classified:

$$s = (1/n)$$

In procedure 2, a third category called undecided was used in addition to the other two. For the procedure

$$s = \{1/[n + (u/2)]\}$$

where u is the number of alternatives in item j that the judge couldn't classify.

The mean and median passing scores were almost identical under the two conditions and both within 2 points of the passing score used by the instructor. The variation for passing scores set by the judges seemed very large (s=6,

no. of items=40, mean=30). The interquartile range was about 10 points.

Poggi, Glasnapp and Eros (1981) have to-date conducted the largest reported comparative study in standard setting methods. They used over nine hundred teachers in a state-wide sample in Kansas with the Nedelsky, Ebel, Angoff and contrasting groups methods, to determine performance in two subject areas (mathematics, and reading) and five grade levels (2, 4, 6, 8, and 11).

In this study, the Ebel method produced the highest cut-off-score, followed by the Angoff and Nedelsky procedures. The contrasting groups procedure fell between the Nedelsky and the Angoff cut-off scores. The difference in average failure rates across all tests ranged from a low of 6.11% between contrasting groups and Nedelsky procedure to a high of 22.49% between Ebel and Nedelsky methods.

Halpin, Sigmon & Halpin (1983) compared the Ebel, Nedelsky, and the Angoff approaches to determine the cut-off scores on the Missouri College English test for admission to teacher education. Fifteen judges divided into three groups (advanced Ph. D. English education teachers; High School teachers and university faculty members) rated a 90-item multiple choice test. Spelling, capitalization, punctuation, sentence style and structure, procedure again produced significantly lower mean cutting score than either the Angoff or the Ebel methods. The Angoff and the Ebel

comparison did not reach significance.

Mills (1983) compared three methods (Angoff, Borderline and Contrasting groups) procedures of establishing cut-off scores on the Louisiana Department of Education second grade basic skills test in Language Arts and Mathematics on items field tested in April 1981. The Angoff and contrasting groups produced similar standards. The borderline cut-off scores were not similar with the other methods.

Seven comparative studies have been cited in which two or more of the popularly used standard setting methods have been investigated. While the results of each of the findings continue to be mixed, there is some consistency to the extent that each method produced consistent information as far as deriving cut-off scores are concerned. In all of the comparative studies cited, the Nedelsky procedure produced the lowest cut-off score, while the Ebel method produced the highest. Other methods were sandwiched between these two extremes.

The consistency of this information seems to convey the notion that the various methods can be used by the various publics for the purpose to which the standard setting exercise is to serve. That is, if the exercise is for graduation, promotion, selection, remediation, job application or certification, then the number one wishes to screen out would be linked to the procedure used to set the pass-fail cut-off.

These studies further point up that if a centrally administered school district wishes to control for comparability of standards among and between its various jurisdictions, or school boards,then the decision must be taken to use the same method or procedure and not allow each jurisdiction or board to choose its own. The importance of this is grounded in the fact that each method produces a different cut-off score or standard and that differ ent raters bring their own levels of expectation to bear on what will ultimately be transformed as a standard. Therefore where education is centrally controlled, the standard setting should reflect a national exercise thus allowing the score to anchor the standards for comparison.

It also seems clear from the results of these studies that little further information could be determined by continuing to engage in empirical comparative investigation. It is more beneficial to now concentrate on the individual studies to firm up judges' competency in dealing with the various methods so that stable scores across and between judges could result. Furthermore, the move now is to incorporate normative test statistics as supplemental information in order to aid judges in their exercise . However, not many studies have been done using the test statistics as supplemental data. This researcher is aware of only two stuies that have utilized the supplemental data as part of the standard setting process, they are: The Jaeger (1982) iterative structure judgement process. In this

process, after the first rating session, judges were shown a graphical distribution of the standards recommended by all the judges in their group along with a cumulative distribution of standards recommended by members of their group as well as the item difficulty index for each item. Armed with this information, judges engaged in a second rating session. At the end of the second session, judges were again given a distribution of the recommended standards, this time a revised distribution.

The third session was similar. At the conclusion of this final session, the scores were summed. The results again were mixed. For example, the registered voters standards were much higher than the others; they felt the standards to be far too lenient. The high school teachers wanted the standards on the Reading test to be lowered slightly, while the principals and counselors wanted it raised. As for mathematics, all the judges agreed that the standard should be raised substantially, but disagreed in the amount it should be raised. The results reported favourably on the feasibility this method to set minimum standards for high school compentency in mathematics and reading.

The second study was done by Cross et al(1984)for the Virginia Department of Education, National Teacher Examination in Elementary Education and Mathematics. In this study they compared the effects normative information would have upon standards derived using three methods: Nedelsky,

Angoff and Jaeger. Thirty faculty members each with at least 2 years experience teaching undergraduate subjects in their field were chosen. They came from seven different teacher training institutions. Fifteen judges were assigned to set standard for the elementary education examination. The other 15 were to set standard for the mathematics examination.

The exercise took three sessions. However, the procedure was somewhat different from that of Jaeger. The important difference was that the judges were required to review the tests in a split-half(odd-even)fashion. During the first and third sessions, they reviewed odd numbered items. The second session they reviewed even numbered items. The reason for this modification was an attempt to avoid any 'carry-over' effect that might have resulted from reviewing the same set of items. The judges were also privy to three types of normative information: a percentage of examinees estimated to know the answer to each question during the second session; a cumulative frequency distribution of scores from a try out sample(used evaluate the stringency of the standard they set in the first session). And finally, the judges in the Jaeger method were given the results of the standard set by the other raters and allowed to evaluate the stringency of their standard in comparison to the other judges.

The result revealed significant main effects for methods and sessions but not for subject areas. There was a significant difference in the first session standard from

the other sessions. The difference was not significant between the second and third sessions. It is suggested therefore that the normative data provided during the second session had some effect upon the raters judgments. As was expected, the Nedelsky method produced the lowest standard than the Angoff and the Jaeger methods. It is also noteworthy to point out that the Nedelsky standard in addition to being less stringent, it was also less reliable and correlated less well with item difficulty values and item relevance ratings(p.127). Finally, the authors concluded that generally, the Angoff method more easily accommodated the normative information than both the Nedelsky and the Jaeger methods.

There is a need to investigate further the effect this supplemental information have on judges' opinion, and to examine the implications of discrepancies between the normative data and the derived cut-off score. Since studies are few, additional experiences with this procedure are needed to assess the utility of this perceived benefit.

2.9.1 Summary

This chapter reviewed firstly the debate surrounding the standard setting issue, the consensus being that all such methods are judgemental but by actually engaging in empirical process of determining a cut-off score you would lessen the arbitrariness and capriciousness of the standard.

Secondly, the chapter reviewed several of the more popular methods in use, pointing out to some extent their strengths and weaknesses as well as some guidelines for their implementation.

Thirdly, some of the comparative studies concerned with the various methods were critiqued. These studies seem to indicate consistently that the Nedelsky procedure produces the lowest cut-off score while the Ebel method produces the highest cut-off scores or standard , with the other methods falling between these two extremes.

3. THE ALBERTA ACHIEVEMENT TESTING PROGRAM

During the first half of the 1970s there were widely circulated reports and fears about a decline in educational standards and lack of a common approach in evaluation techniques expressed by the media, parents and universities in Alberta. The Minister's Advisory Committee on Student Achievement (MACOSA) was commissioned in November 1976 to investigate these concerns and to make recommendations for their solutions.

The Alberta Achievement testing program arose from a result of the recommendation of this committee. Amongst its recommendations, the committee suggested that:

1. achievement tests be administered periodically at selected grade levels, but with sampling plans extended to permit generalization of results to a school system population.
2. the proposed assessment program test curricular objectives selected from all domains (knowledge, attitudes and skills) and thought levels. It was further expressed that a set of standards would evolve that could be used for interpreting student achievement in the various jurisdictions.

However, after the results of the initial test, it became clear that in addition to the raw data provided to the jurisdictions, a set of interpretive guidelines was necessary, otherwise the results would be of limited use to educators across the province (Maguire 1982).

3.0.1 Goal of the Alberta Achievement Program

The goal of the Alberta Achievement testing program is to provide feedback. In the Alberta context, it is to provide feedback to jurisdictions, as to whether students in the province are achieving curriculum objectives as well as they should. This feedback can only be meaningful if the interpretation of the level of achievement representing satisfactory attainment of these objectives can be defined. The definition constitutes a standard against which the jurisdictions might interpret student achievement.

3.0.2 Interpretation of Test Scores

Test score interpretation for individuals maybe viewed from two perspectives-norm referencing and criterion referencing. What distinguishes these two types of interpretations from one another is the manner by which meaning is drawn from the scores. In the case of norm referencing, meaning is derived from the relationship of a single examinee to the total population that has been tested. In criterion referencing, the individual's position on the test is not important but rather his or her performance with relation to a specified standard or criterion In other words, norm referencing refers to the derivation of meaning from a person's position relative to other people, and criterion referencing refers to the derivation of meaning from the tasks performed in the evaluation.

Maguire (1982) contends, when dealing with jurisdictions, the same two referential bases could be considered. However, he warns, in the Alberta context, "referring jurisdiction averages to distribution of jurisdiction values presents serious conceptual problems" (p.1). He cites the wide variation in jurisdiction size that exists across the province as the main problem. At most grade levels, the ratio of the size of the largest jurisdictions to the smallest size is greater than 100 to 1. Because of their size, the very large jurisdictions are likely to be located near the centre of the distribution, whereas jurisdiction falling at the extremes will tend to have fewer students. The reason for this being that it is easier to change the scores of 30 people than of 3000 people. In order for a large jurisdiction (for example 3, 000 students) to have an average score near the top of the distribution, it must have reasonably high scores for 3000 students. Thus, it is much more likely that a jurisdiction with only 30 students will fall near the top of the distribution of jurisdiction means (Maguire 1982).

A second problem with norm referencing Maguire's paper highlighted is the fact that it necessarily produces winners and losers. Yet it is conceivable that the state of achievement in some curricular areas may be inadequate in all jurisdictions, or it may be very satisfactory for all jurisdictions in other curricular areas. While both of these situations are plausible neither would be identified under a

norm referenced system.

These two arguments are compelling, yet there are some points in favour of using some aspects of norm referencing. About 25,000 students participate in any test administration. Their results provide a reference level against which the suitability of the test itself can be assessed.

According to Nitko (1980), there are a variety of criterion referenced systems. For the Achievement Testing program, a concerted attempt was made to have items derived from a set of well-defined curriculum specifications. (This corresponds to Nitko's second category, "Criterion-referenced tests based on well-defined but unordered domains"). The curriculum specifications were supplied by the Curriculum Branch of Alberta Education, and in most cases the importance of each objective was included in the specification. From the specifications, test blueprints were prepared and items written.

The exact process of content validation used, differed from test to test, but in all cases technical review committees examined and approved the final selection items. After the tests had been administered the same committees assisted in preparing the reports to be presented to the various audiences. Maguire (1982) asserts, "in some cases implicit interpretations of the results were made, but for the most part the reports are characterized by description rather than judgment. The prevailing belief appeared to be

that the meaning of results derived directly from the kinds of tasks that the children could perform, and in that respect the referential basis was criterion referencing" (p.3).

The problem with the process, Maguire warns, was that it did not provide any standards against which provincial or jurisdictional performances could be measured. Consequently, the Student Evaluation Branch initiated the move toward using a variety of these methods in its quest to provide reference levels against which to make judgements of student performance. As has been pointed out elsewhere in this study, the Student Evaluation addressed that problem by testing the use of standard setting methods in its achievement testing evaluation. One of the methods used was The Nedelsky's (1954) procedure.

3.0.3 Summary

This chapter discussed the genesis of the standard setting momentum in the Alberta Education department. It revealed that the impetus was spawned by the MACOSA commitee's recommendation. The problem of interpreting test results for the jurisdictions was also brought into focus and discussed. The salient point was the need to establish a referential level for interpretation, and this could best be done by using judgmental standard setting methods as opposed to the outright norm referencing that heretofore was the procedure.

4. METHODS AND PROCEDURES

This chapter presents a description of the instrument used in the study, as well as a description of the method and procedure used in data collection and analysis for deriving a cut-off score on the multiple-choice section of the Alberta grade 9 Social Studies achievement test.

The data for the present study including teacher judgments and student responses were collected by the Student Evaluation Branch as a part of their in house research and development program. The reanalysis of this information is the focus of the present study.

4.0.1 Instrument

The test used for this study was the 1983 Alberta Achievement grade 9 Social Studies. The test was designed to provide extensive and reliable information as to the level of student achievement in basic education throughout the Province.

The grade 9 Social Studies achievement test consisted of two sections: a multiple choice section weighted 70% and a written response section weighted 30%. For the purpose of this study, the multiple choice section only, was used. This section consisted of 60 items designed to reflect the values, knowledge and inquiring skills as outlined in the curriculum specifications.

4.0.2 Method and Procedure

During June 1983 the Province of Alberta administered the grade 9 Social Studies test. A total of 23, 579 students in 564 schools from 104 jurisdictions wrote this examination. The test booklets were returned to the Alberta Student Evaluation Branch(SEB) for scoring and determining a minimum cut-off score for the test.

A sample of 84 teachers was used as raters in the standard setting and marking exercises. Each teacher (rater) had to be in possession of a valid Alberta teachers certificate, have taught grade 9 Social Studies for at least three years and be currently teaching grade 9 Social Studies. The multiple choice section of the test was machine scored and analysed. However, teachers used the multiple choice items initially without the data analysis to determine a miminum cut-off score that would be acceptable as a pass in the province.

A modification of the Nedelsky (1954) procedure was employed to determine the standard. Raters were engaged in two sessions. Each rater was required to do two tasks:

1. To rate each of the 60 multiple choice items in terms of its curricular relevance as essential (E), important (I), adequate (A) and questionable (Q).
2. To circle a number of alternative(s) from 0 to 3, to indicate how many of the distractors a 'borderline student' could eliminate. For the purpose of this study a 'borderline student' was considered to be one at the fifteenth

percentile. When translated to an "average" of 30 students, this student would be third from the bottom; the assumption being that there exists an average distribution of ability in the class.

After each rating session, each judge was given the actual item statistical analysis for 600 students as well as their derived borderline score. They were then asked to repeat part two of the task. They could if they wished, revise judgments after being exposed to the actual test data and if they thought it necessary.

Upon completion of the second session, the overall standard was obtained by summing the item scores of each judge and these sums were then averaged across judges. This standard also represented the total test score expected from the referenced group.

4.0.3 Data collection

Data for this study came from the marking (scoring) sessions conducted by the Student Evaluation Branch of the Alberta Education. The item analysis from the social studies test was also obtained from the Student Evaluation Branch.

Eight-four teachers were initially engaged in the standard setting and scoring exercies, selected from 104 jurisdictions. However, all 84 teachers did not complete the two sessions. Those teachers not completing both sessions were subsequently removed from the final list. Having removed those, a total of 71 judges participated in the

completion of the standard setting session.

4.0.4 Treatment of Data

The information was summarized using exploratory data analysis along with descriptive statistics.

The data from the teachers consisted of two ratings for each item. The first rating was the importance on the four point scale, and the second was a number from 0 to 3 indicating the number of distractors that a borderline student could eliminate. The test data were the responses made by the 23, 579 students on the 60 items. Conventional item analysis gave item difficulties, item test correlations, frequency distribution of raw scores and test reliability. For each judge a cut-off score was found using Nedelsky's procedure. An item passing level is defined as 1/(4-I) where I (the item judgement) is the number of alternatives which that judge estimated a borderline student could eliminate. The item passing levels were then averaged over items for each judge, to get the judge's standard. Finally these standards were averaged over judges to get an overall standard. This standard is the proportion of items that borderline students should be able to pass.

To determine if judges using the same method arrived at the same or different standards, the judge's standard was multiplied by 60 to give a value for Time1 which was the test score which that judge designated as the standard at Time 1. Similar scores were calculated for Time 2. The

difference between Time1 and Time2 is an indication of the difference in standards used by the judges.

Frequency counts for Time1 and Time2 were derived and were compared with the scores of the observed test score distribution at the 15th and 16th percentile which was the derived borderline value set in the original test by the SEB.

To ascertain whether some judges changed more often than others in their ratings of the various items, T1 , (the item judgement at Time 1) was subtracted from T2 for each item. For each judge the differences between item judgement were tabled to see how often they made big changes, small changes or no change. This process was collapsed over all judges to get the overall effect.

To measure the amount of change in an item, the judge ratings T2 - T1 were examined. A positive shift indicated a change toward rating the item as harder, while a negative shift meant that the judge rated the item as easier following the receipt of information.

It was further the purpose of the study to determine if judges after classifiying items as essential were more or less resistant to change than if they classify items as Acceptable (A), Essential (E) or Questionable (Q). For this a frequency count of the items classified, Acceptable, Essential and Questionable was obtained along with their corresponding judgement to indicate this resistance. A chi square was computed to ascertain this relationship.

To find out if raters' judgments of EIAQ classification related to item statistics, a frequency of each item so classified was computed along with each corresponding item statistic for the comparability for each judge rating and as a group.

A contingency coefficient, was used to compute each judge rating for each item classified as E, I, A, Q, in comparison with the actual test statistics of item difficulty.

5. RESULTS AND DISCUSSION

The results of the analysis of the data are presented and discussed in this chapter in the order of the questions as stated in the study.

The average of the final rating recommended cut-off score for the multiple choice section of the grade 9 Social Studies Achievement Test by 71 teachers (judges) was 22.526 items for a borderline student out of a 60 item test. In other words, the judges recommended that a borderline student must be able to get a minimum of 23 items (rounded to the nearest whole) correct in order to pass the Alberta grade 9 Social Studies Achievement Test.

The remainder of the results will be discussed question by question.

Question One:

Do judges using the same test content and the same method set the same or different standards?

Table 3 shows a stem and leaf display of the judges recommended cut-off score (rounded to the nearest score) for the first rating session. The scores are very dispersed with a range of 18 to 46 items, with a median of 28 and an interquartile range of 10. One half of the cut-off scores fell between 25 and 35. There were no outliers. It is noted that the recommended scores among judges are extremely variable. For example, one judge recommended a cut-off score of 18 items, while 16 judges recommended 28 or 29 items and another judge recommended 46 items to be the cut-off point

in the 60 item test for borderline student success.

Table 3: Stem and Leaf display of Judges Standard Scores for Time 1 Rounded to Nearest Whole Number

	f	
1		
* 8	1	
2- 0111	4	
t 23	2	
f 444555555555	12	median=28
s 6666777	7	Q_1 25 35=Q_3
* 8888888888999999	16	18 46
3- 00000	5	
t 22	2	Range=28
f 44455555555	11	Interquartile Range=10
s 66777	5	
* 888	3	No Outliers
4- 1	1	
t	0	
f 4	1	
s 6	1	

The wide spread in individual judges' recommended cut-off scores substantiate the findings of previous studies (Andrew & Hecht, 1976; Koffler, 1980; and Skakun & Kling, 1980) that judges do not set the same standards regardless of their similarity in backgrounds. This notion also seems to find support in Glass' argument that it is difficult if not impossible to set standards on the definition of minimal competence or borderline in any unequivocal or psychologically meaningful way to the point that all judges would have a clear representation of a borderline student for whom they were to set a criterion measure.

It was perhaps not expected that judges would set the same standards, but rather that they would set similar standards, bearing in mind that they were coming from similar professional and experiential backgrounds. Thus, the

wide variability in their judgments gives reason for some apprehension. However, Popham (1981) attempts to allay this concern when he notes that:

> When human beings apply their judgmental powers to the solution of problems, mistakes will be made. However, the fact that judgmental errors are possible should not send us running from such tasks as setting standards. Judges and juries are capable or errors, yet they do the best they can. Similarly, educators are now faced with the necessity of establishing performance standards, and they too must do the best job they can (p. 374).

The tenor of this is reflected in the need for adequate training and understanding in and of the methods. The lack of this training and understanding of the task may have been one of the reasons for the wide variability among judges' scores. With adequate training Popham (1981) asserted that "the act of standard setting can reflect the very finest form of judgmental decision making."

Question Two:

> What effect do test statistics data, when used in an iterative standard setting context, have on influencing judges' opinions?

The results in Table 4a show a stem and leaf display of the judges' recommended change cut-off scores (rounded to the nearest scores) after the second rating session when judges

would have been given access to actual test statistics. The recommended cut-off scores on this second rating session was a median of 22 and a narrower range of 22 items with a low of 18 and a high of 40. The interquartile range of 4 was much smaller than the value of 10 obtained at time1. The scores, though variable, were less variable than the initial judgments. Thirty-four judges recommended cut-off scores below 22 while 25 were above the score of 22. It is also worth noting that there were 4 outliers. That is to say, 4 scores were more than one and a half interquartile range above the third quartile.

Table 4a: Stem and Leaf Display Judges' Recommended Cutoff Scores Time 2 (Rounded to Nearest Score)

1-		0	
s		0	
*	88888888999999	14	
2-	00000000111111111111	20	
t	222222222222333333	18	median 22
f	4444555	7	Q_1=20 Q_3=24
s	7777	4	
*	8899	4	Low=18 40=High
3-	1	1	
t	2	1	Range=22
f	4	1	Interquartile Range=4
s		0	
*		0	Outliers 31, 32, 34, 40
4-	0	1	

In table 4b the time2 results have been clustered. An attempt was made to have one fifth of the judges in each group. It is noted that the lowest group of 14 judges recommended cut off scores of 18 and 19 items, while the top 15 judges recommended scores from 25 to 40 items, indicating that the distribution of recommended scores is positively

skewed.

Table 4b: Distribution of Judges Standard Scores for Second Rating

Standard Score Interval	No. of Judges	%	Mean	SD
17.67-19.33	14	19.7		
19.50-20.92	12	16.9		
21.00-22.08	15	21.1		
22.17-24.33	15	21.2		
24.50-39.83	15	21.1		
	N=71	100	22.566	4.049

Table 5: Mean Comparisons Between Judges Standard Score for Time 1 and Time 2 Ratings

	Time 1	Time 2	t	p
Mean	29.53	22.57	15.13	0.00
SD	5.72	4.049		

df = 70
r = .74($p<.001$)
N = 71

A summary of the mean comparison of judges' ratings for both sessions is provided in Table 5. For the first session judges were required to indicate item by item the number of alternatives in a four option item that a borderline student would be able to recognize as being incorrect and thereby eliminate. At the end of this session, they were provided with actual test data distribution of multiple choice scores for a sample of 600 students and the borderline cut-off score derived from their first judgments. Judges were then asked to repeat the task a second time, thus being given the opportunity to change their judgements if they thought it

necessary, based on the additional supplemental information. The results in table 5 show a mean comparison of judges' ratings for both sessions. A t-test was performed. The results indicate a significant difference ($p<.001$) between the average recommended cut-off score by judges' first ratings and the second. Table 6 provides a visual back-to-back stem and leaf display of the change that occurred between the first judgment and the second. It is noted from the table that whereas the first ratings were very spread and multimodal with most of the scores falling between 24 and 34 items the distribution of the second rating had changed substantially. The borderline cut-off scores had a much smaller range of 22, and a smooth distribution, as opposed to their initial rating which was multimodal and very variable. There were four upper outliers in the time2 distribution(outliers refer to judges whose score fell outside q1 and q3). The indication is that a common understanding among most judges may have been reached.

It is perhaps not surprising that the overall opinion of judges would have been influenced after being privy to additional normative information, then repeating the task. The intent of such normative information was to assist judges in reflecting upon and revising if they so wish, their initial judgments in light of new information. It should be pointed out that although the judges at the second rating session were aware of the actual distribution of

Table 6: Back-to-Back Stem and Leaf Display of Judges Standard Score for Time 1 and Time 2 (Rounded to Nearest Score)

f	Time 1		Time 2	f
0		1-		0
0		s		0
1	8	*	88888888999999	14
4	1110	2	00000000111111111111	20
2	32	t	222222222222333333	18
12	555555555444	f	4444555	7
7	7776666	s	7777	4
15	9999998888888888	*	8899	4
5	00000	3-	1	1
2	22	t	2	1
11	55555555444	f	4	1
5	77766	s		0
3	888	*		0
1	1	4-	0	1
0		t		0
1	4	f		0
1	6	s		0
N=71				N=71

Median	28	22
Q_1	25	20
Q_3	35	24
Range	28	22
Interquartile Range	10	4

multiple choice scores, it would have been difficult for any individual judge to attempt to set the standard at a specific multiple-choice score, as the only input was at the individual item level, and the judgments had to be made in relatively crude terms.

It is worth pointing out too, that some judges were not influenced by the additional information. For example, one judge had recommended a cut-off score of 18 items on the first rating and again on the second rating. He may have changed on the individual items, but the cummulative effect

remained the same. The point, however, is further elucidated in the third question.

Question Three

> Do some judges change more than others after receiving the normative feedback ?

Table 7 presents a scatterplot of individual judgment change on cut-off scores from time 1 against time 2. The scatterplot shows a positive correlation ($r=.74$, $p < .001$) between judgments.

Table 8 also shows a summary of judges change pattern overall, as well as the individual change patterns. Time1 scores have been clustered in 5 groups so that approximately one fifth of the judges were in each group. The same categories were used to group the scores for time2. The pattern that emerges from this table indicates two kinds of change—a general group change whereby the majority of teachers engaged in the exercise dropped their initial recommended cut-off score for the second rating, and an individual movement whereby they either shifted substantially, moderately, or not at all. For example, 9 persons who on the first rating had recommended that a borderline student get between 35 and 46 items did revise their judgments to cut-off scores ranging between 10 and 25 items. Sixteen judges did not shift their judgments one way or another. Two of these judges had recommended cut-off scores well over the median score for the group. They did

Table 7 Scatterplot of Judges Scores Time 1 vs. Time 2

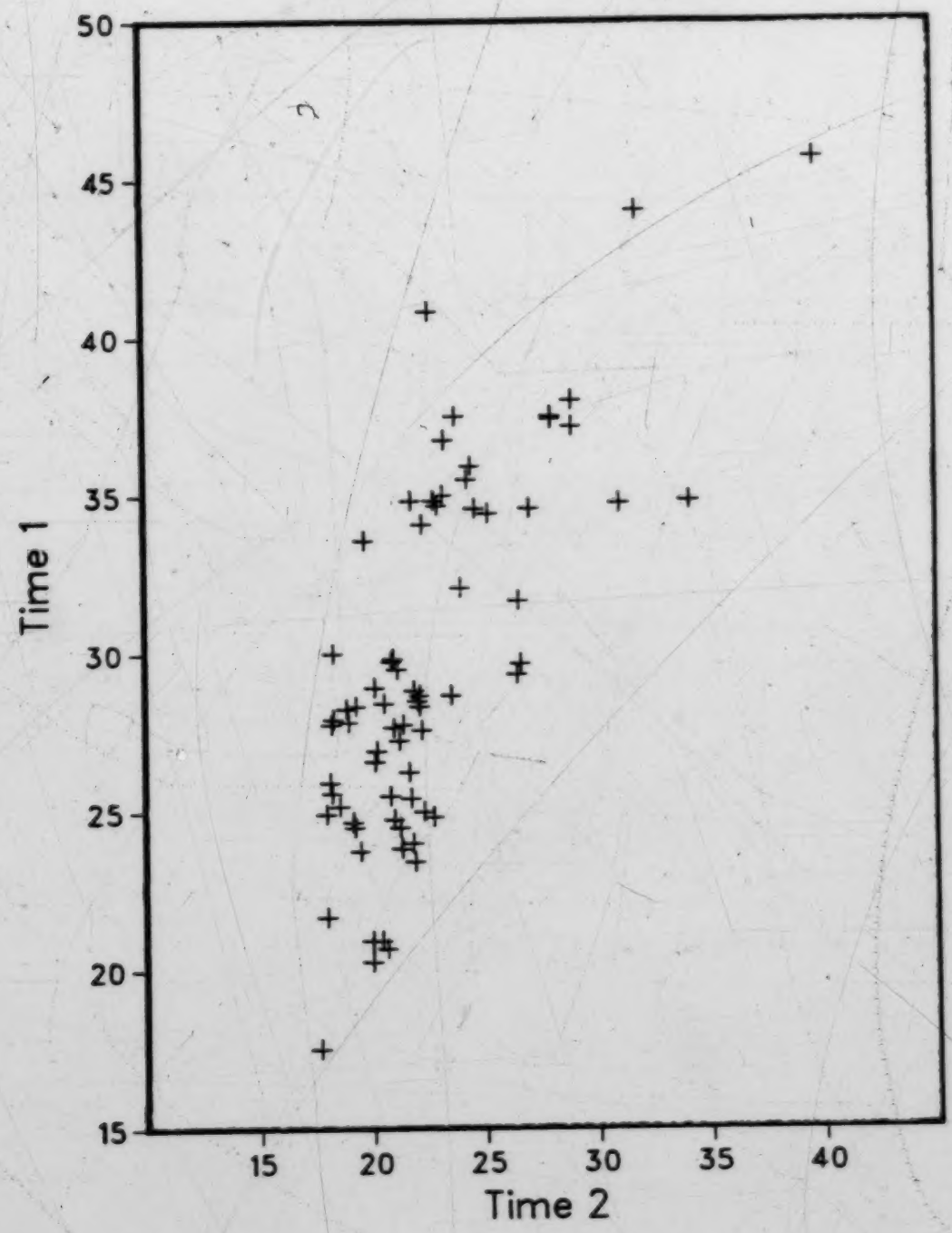

Table 8: Summary of Judges Overall Level and Individual Change From Time 1 vs Time 2

		Time 2					
		1 10-24.7	2 24.8-27.6	3 27.7-29.4	4 29.5-34.7	5 34.8-45.6	
5	34.8-45.6	9		4	2	1	16
4	29.5-34.7	8	4		1		13
Time 1 3	27.7-29.4	13	1				14
2	24.8-27.6	14					14
1	10-24.7	14					14
		58	5	4	3	1	

not changed their opinion at the second rating. It ought to be noted too that whereas 14 judges began by setting a cut-off score between 17.5 and 24.7 items, after the second rating 81.6% or 58 judges had changed their cut-off score to that interval. In other words, the majority of the judges had changed by shifting their initial judgments downwards.

Table 9: General Group Changes

	# of Items Changed	Number of Judges
High Changers	37 or more	21
Medium Changers	9 to 36	30
Low Changers	8 items or less	20

The researcher has alluded earlier to the general group change. Table 9 shows the three distinct groupings: Those judges who changed many items, or high changers, those who changed only a few items, or low changers, and those who were moderate or medium changers. For the purpose of this study a high change judge was defined as one who changed the rating of 37 items or more, while a member of the low group was one who changed ratings on 8 items or less. The medium group fitted on a continuum between these two extremes. It is construed from Table 9 that the high change group was apparently the group most influenced by the additional normative information, while the low group is seen as being least influenced. Perhaps one possible explanation for 16 judges not to be influenced by the normative feed-back might be associated with the nature of the Nedelsky method itself.

In a similar study by Cross et al(1984)they noted that several judges "expressed difficulty in judging which distractor a minimally competent individual would recognized as incorrect"(p.126). As a result they observed that the the Nedelsky method, after the introduction of the normative feed-back were less affected when compared with the the other methods. They surmized that ".....*some* reviewers were less sure about how to incorporate item difficulty information into their judgments"(p.126). This apparent confusion may quite conceivebly explain why the 16 judges in the present study were not affected by the normative data.

As an aside to the question, it is worth mentioning that there were 3 items (4, 7, and 22) in the test that appear to have been significantly affected by the teachers' judgments after the second rating. Table 10 presents a modal distribution of the judges' estimated item difficulty for both occasions. For item 7, for example, at time 1, judges indicated that all the alternatives should be eliminated by the student. In other words, the item was seen as very easy. However, after the time 2 rating, judges had changed their opinion that a borderline student could not reject any of the distractors. It turned out that the judges' assessment of the item after the second rating was fairly accurate. Appendix B indicates that 39% of the students (which would likely include all of the borderline students) failed that item. Another interesting item was item 45. Again the judges felt initially that it was moderately easy, but after the

Table 10: Modal Summary of Judges Item Ratings Difficulty and Shifts for Both Sessions

		Second Rating by Item and Difficulty				
		0	1	2	3	
First Rating by Item and Difficulty	0	28, 43, 49	—	—	—	3
	1	5, 34, 40, 53, 56, 60	2, 6, 12, 16, 19, 24, 30, 33, 37, 42, 44, 57	—	—	18
	2	20, 35, 45	1, 3, 8, 9, 10, 11, 14, 15, 17, 18, 25, 26, 27, 29, 31, 36, 38, 41, 46, 47, 48, 50, 52, 54, 55, 58, 59	39, 51	—	32
	3	7	4, 22	13, 21, 23, 32	—	7
	Σ	13	41	6	0	60

second rating revised their judgments to view it as a very difficult item for borderline students. Once again the actual item statistics found in Appendix B indicates they were correct in the judgments, as 51% of all the students got it wrong. These observations of the judges' tasks is simply to reinforce Shepherd's point, that judges would make more reasonable and informed judgments if they are provided with normative data and if the standard setting process is an iterative one.

Question Four

> If judges classify items as essential, are they more or less resistant to change than if they classify the item as important, acceptable or questionable?

For the purpose of this study, Time 2 minus Time 1 ($T_2 - T_1$) overall item judgment collapsed over all 71 judges was taken as a measure of the amount of change on a scale ±3 points.

Table 11 presents a relationship between item relevance and judgment change. A chi square test was performed between judges' estimated item difficulty and item relevance and judgment change. The results indicate a significant relationship, chi square= 35.036, $p<.001$). That is to say, that there seems to be a relationship between judges' perception of the item relevance and the estimated item difficulty. The chi square contribution in each cell confirms the previous observation that the items classified as questionable have a different pattern than the others.

Table 11: Chi Square Relationship Between Judges Estimated Rating of Item Difficulty to Item Relevance

Degree of Change	RELEVANCE A	(O-E)	X^2	B	(O-E)	X^2	C	(O-E)	X^2	D	(O-E)	X^2
-2	97	-10.71	1.064	211	17.12	1.511	178	-3.67	0.164	13	-2.73	0.473
-1	435	36.82	3.404	718	1.28	0.002	288	-13.92	0.641	34	-24.16	10.036
0	588	-29.92	1.448	1098	-14.26	0.182	490	24.46	0.482	113	22.73	5.723
1	28	3.17	0.404	42	-2.74	0.162	14	-4.88	1.251	8	4.38	5.299
2	2	0.66	0.325	1	-1.42	0.833	2	0.98	0.941	0	-0.19	0.19

$X^2(12) = 35.036$, $P < .001$

Note: (-) made item more difficult
(+) made item easier
(0) no change on item difficulty by judges

A = Essential Items
B = Important Items
C = Acceptable Items
D = Questionable Items

Table 12: Distribution of Judges rating of Item Relevance to Degree of Estimated Change

Direction of Change	Relevance E	I	A	Q	Total	
-2	97	211	178	13	399	
	(.08)	(.10)	(.09)	(.07)		44%
-1	435	718	288	34	1475	
	(.38)	(.34)	(.33)	(.20)		
0	588	1098	490	113	2289	54%
	(.51)	(.53)	(.56)	(.67)		
1	28	42	14	8	92	
	(.024)	(.020)	(.016)	(.047)		2%
2	2	1	2	-	5	
	(.001)	(.0004)	(.002)	-		
Total	1150	2070	872	168	4260	

Note: (-) made item more difficult
(+) made item easier
(0) no change on item difficulty by judges

4260 = 60 items x 71 judges

Table 12 reveals that there was not much difference between judges' change pattern on items classified as essential or important. For example, 49% of the 1150 item judgments classed as essential were changed while 47% of the 2070 item judgments on important items were changed. The same pattern is true for items classified as acceptable. Out of 872 items so classified, 44% or 382 were changed. The same pattern does not emerge for items classed as questionable, however, Only 33% of the 168 item judgments in this category were changed. Thus, the judges seem to be more flexible on the items they consider important than the items they considered questionable. Perhaps, judges were less concerned with items classified as questionable. Similar results were found in

the Cross, Impara, Frary & Jaeger (1984) study.

Contingency coefficients were calculated for each item between judged relevence and change in judged difficulty. Tables 13 and 14 show these coefficients for all the items. The median estimated contingency coefficient was .36, at Time1 which suggest that a moderate relationship existed at time1.

Table 14, indicates that the item judgments after the second rating were slightly less related than they were at time1 The median coefficient had shifted to .30. It is to be noted from both tables however, that for some items there is a big relationship, while for others the relationship is small. Similarly, for some items that showed a big relationship at time1, showed only a small to moderate relationship at time2.(For example, items 47 and 31 indicated strong relationship of item relevance to Judges estimated difficulty at time1).But those items indicated substantially more moderate relationship at time2. Likewise, item 23 indicated moderate relationship at time1 but a very strong relationship at time2. The reverse of this is seen with item 21. The initial relationship was very weak, but time2 rating showed that it was slightly above the median coefficient at time2. These tables suggest that the reasons for changing ratings were not consistently associated with perceptions of importance. Perhaps the individual nature of the items e.g. specific content had more to do with influencing judges behavior.

Table 13: Histogram of contingency Coefficient of the Relationship of Items Relevance to Judges Estimated Difficulty

Contingency Coefficient	First Rating Item Numbers	Frequency
.0-.05		
.06-.10	21 22	2
.11-.15		2
.16-.20	1 30 34	3
.21-.25	2 5 9 12 15	5
.26-.30	4 7 17 18 45 56 59	7
.31-.35	3 6 8 15 19 32 35 57 58 40 49 57	12
.36-.40	13 16 20 23 27 28 36 39 41 44 46 48 50 55 58 60	16
.41-.45	10 24 26 29 42 43 51 53	8
.46-.50	11 14 33 52 54	5
.51-.55	31	1
.56-.60		0
.61-.65	47	1
		--
		60

Median .36 Q_1 = .290 Q_3 = .405

Table 14: Histogram of Contingency Coefficient of the Relationship of Item Relevance to Estimated Difficulty

Contingency Coefficient	Second Rating Item Numbers	Frequency
.0-.05		
.06-.10		0
.11-.15	6	1
.16-.20	7 11 22 37 55	5
.21-.25	1 18 19 26 30 51 60	7
.26-.30	2 4 5 8 10 17 20 24 25 28 31 34 35 41 43 44 48 59	18
.31-.35	9 12 15 21 39 40 42 45 50 53 54 56 57	13
.36-.40	27 29 32 33 36 38 46 49 52 58	10
.41-.45	3 13 14 16 47	5
.46-.50		0
.51-.55		0
.56-.60	23	1
		--
		60

Q_1 = .31 and Q_3 = .36
Median .30

Question Five

How do judges' estimates of item difficulty compare with the actual test item difficulty?

One of the tasks the judges had to perform was deciding item by item the number of alternatives for each item a typical 15th percentile student would eliminate as obviously wrong. The number of distractors eliminated, then, is linked in this case, at least theoretically, with the P-value (difficulty level) of the actual test item statistics. For instance, if the judge thought a borderline student could not reject any alternatives for an item, that item received an estimated difficulty rating of .250 on the assumption that a borderline student would choose at random from among the four alternatives and would therefore have one chance in four in getting the item correct. That is to say, the item was perceived to be a difficult item by that judge. Similarly, if a judge felt that a borderline student could eliminate three distractors for an item, that item was assigned an estimated difficulty of 1.000, which is to say that the item was judged to be extremely easy. Each judge's estimated item difficulty for each item was collapsed across all judges to arrive at the overall estimated difficulty for each item. The actual test item difficulty was computed by means of conventional item analysis as the percentage of students who correctly answered an item. The judges' estimated judgment of item difficulty was compared with the

actual test item difficulty. Table 15 shows the relationship between the judges' estimated item-by-item difficulty for the first rating session with the actual test item difficulty. The estimated item difficulty was classified according to the modal ratings of the judges. The results indicate that the majority of the items estimated by the judges to be medium to hard were easy to medium on the actual test statistics. The correlation turned out to be r=.33.

Table 16 gives a cross tabulation of judgments and empirical difficulties. Again it can be seen that judges had estimated 45 items as medium to hard (0-2) whereas the actual item difficulties were pegged easy to medium.However, it should be remembered that the judges were thinking about borderline students in this exercise. A chi square of this relationship was performed. The result is noted in Table 17, and indicates significance at the .001 level ($p<.001$). However, after the second rating which was then taken as the final results for the exercise, judgments were seen as less conforming to the actual test statistics. Table 18 shows quite a heavy concentration of judges' estimations of items to be more difficult than they thought at the first rating. Table 19 indicates that judges after the second session rated 51 items to be either medium or hard. These same items were assessed using conventional item analysis to be of easy to medium difficulty. This second session's judges' estimated item difficulty correlated very low (r=.14) with

Table 15: Relationship of Judges First Ratings of Item Difficulty to Observed Test Item Difficulty

Judges Estimate of Item Difficulty Time 1		Observed Item Difficulty				
			Hard	Medium	Easy	
		.0-.20	.21-.45	.46-.60	.61-.80	.81-1.00
Hard	0	—	49	—	43	
	1	—	12, 24, 30, 34, 37, 60	16, 40, 53, 57	5, 19, 28, 33, 42, 44, 56	—
Medium	2	—	2, 3, 26	9, 14, 17, 18, 25, 29, 31, 39, 41, 45, 46, 51, 54, 59	1, 6, 8, 11, 15, 20, 27, 35, 36, 38, 47, 48, 50, 52, 55, 58	10
Easy	3	—	—	—	7, 13, 21, 22, 32	4, 23

$r = .33$

Table 16: Frequency Distribution of Relationship of Judges First Estimated Ratings of Item Difficulty to Corresponding Observed Test Item Difficulty

		Observed Test Item Difficulty			
		First Rating			
Judges' Estimate of Item Difficulty		HARD	MEDIUM	EASY	
		.0-.45	.46-.60	.61-1.00	
HARD	0/1	7	4	9	20
MEDIUM	2	3	14	18	35
EASY	3	-	-	5	5
		10	18	32	60

r=.33

the actual item difficulty levels of the test. The chi square computed failed to reach significance ($p>.05$). This result is consistent with the observation that the judges tended to move all items into the medium and hard category. One possible explanation for this might be that judges initially over estimated the capabilities of the borderline students. The fact that there was no significant difference at time2 seem to suggest that the information provided as supplemental or reality check data probably did have some general influence on some of the judges.

Question Six

If judges do change, do they tend to change towards congruence? That is, do they change towards making the items more like the observed difficulty or less like the observed item difficulty?

The results shown in Tables 16 and 18, which represent the number of items changed overall by the group, suggest that judges do change their judgments between rating sessions. Table 16 shows a correlation, albeit low, ($r=.33$) between judgments of estimated item difficulty to actual test statistics; which is to say, the first rating was somewhat like the actual test statistic. This is further supported in Table 17, which shows a chi square of significance ($p<.001$) between judges' estimate of item difficulty to actual observed item difficulty.

When we look at Tables 18, 19 and 20, which are the computations of the second ratings, it is seen that not only did judges change, but they did so while making the items more unlike the observed difficulty. The correlation as shown in Table 19 is very low ($r=.14$). The chi square test was performed and is seen in Table 20. This failed to reach significance ($p>.05$). It is to be taken therefore that when judges did change, they changed towards making the item less like the observed item difficulty. This was perhaps to be expected in the sense that the judges' task was to set a minimum cut-off score for a borderline student. Were the judges to have changed to making the items more like the observed item difficulty would have defeated the purpose, since the criterion cut-off score would have very much resembled the distribution of the scores derived via the classical norm referenced approach.

Table 17: Chi Square Relationship of Judges Estimated Item Difficulty Ratings to Observed Test Item Difficulty for First Rating

	.0-.45	(O-E)	$\frac{(O-E)^2}{E}$	.46-.61	(O-E)	$\frac{(O-E)^2}{E}$	.61-1.00	(O-E)	$\frac{(O-E)^2}{E}$
01	7	3.667	4.034	4	-2.00	0.667	9	-1.667	0.260
2	3	-2.833	1.375	14	3.5	1.167	18	-0.667	0.023
3	—	-0.833	0.833	—	-1.5	1.5	5	2.333	2.040

$X^2(4, N=60) = 20.52, P < .001$

Table 18: Relationship of Judges Second Ratings of Item Difficulty to Observed Items

Judges Estimate of Item Difficulty Time 2		Observed Item Difficulty				
		Hard		Medium	Easy	
		.0-.20	.21-.45	.46-.60	.61-.80	.81-1.00
	0	—	34, 49, 60	40, 45, 53	5, 7, 20, 28, 35, 43, 56	—
Hard	1	—	2, 3, 12, 24, 30, 37	9, 14, 16, 17, 18, 25, 26, 29, 31, 38, 41, 46, 54, 57, 59	1, 6, 8, 11, 15, 19, 22, 27, 33, 36, 42, 44, 47, 48, 50, 52, 55, 58	4, 10
Medium	2	—	39, 51	—	13, 21, 32	23
Easy	3	—	—	—	—	—

r = .14

Note: rating of 0 = .250
1 = .333
2 = .500
3 = 1.000
represent the number of distractors a borderline student might eliminate and hence the difficulty of the items.

It appeared, therefore, that the judges used the normative data very well in attempting to lower the cut-off score for borderline students to experience success on the achievement test.

Table 19: Frequency Distribution of Relationship of Judges Second Estimated Ratings of Items to Observed Test Item Difficulty

		Observed Test Item Difficulty			
		Second Rating			
Judges' Estimate of Item Difficulty		HARD .0-.45	MEDIUM .46-60	EASY .61-1.00	
HARD	0/1	10	16	28	54
MEDIUM	2	-	2	4	6
EASY	3	-	-	-	-
		10	18	32	60

Table 20: Chi Square Relationship of Judges Estimated Item Difficulty Rating to Observed Test Item Difficulty for Second Rating

	.0-.45	(O-E)	$\frac{(O-E)^2}{E}$	.46-.60	(O-E)	$\frac{(O-E)^2}{E}$	.61-1.00	(O-E)	$\frac{(O-E)^2}{E}$
01	9	0.9	0.1	18	—	0.000	27	0.0	0.029
2	—	-0.9	0.9	2	—	0.000	4	0.9	0.216
3	—	0.0	0.00	—	—	0.000	—	—	0.000

$X^2(4, N=60) = 9.245$, $P > .05$
*X^2 did not reach significance.

6. SUMMARY, CONCLUSIONS, IMPLICATIONS AND RECOMMENDATIONS

This chapter provides a brief summary of the literature pertaining to the standard setting methods movement as well as a summary of the results to the research questions. Secondly, it gives the conclusions drawn from the results of the study, and lastly, it gives recommendations concerning research findings with several new questions for possible further investigation.

The practice of setting performance standards or cut-off scores on educational tests gained momentum at the turn of the 1960's. Two factors influenced the acceleration of this practice. First was the general trend toward local autonomy in education which meant that classroom teachers were required to set standards at the classroom level without the benefit of knowledge of the performance of the wider population of students. Second was the influence of the objective-based testing movement which urged the abandonment of setting standards by comparisons with other students and adoption of a more 'absolute' approach.

Several methods to set 'absolute' standards were developed. These various methods, as pointed out in the study were classified into three models: Judgmental methods, Empirical methods and Combination methods. Each model has a number of methods with their own peculiar properties, description and underlying assumptions. These methods permeated the measurement community with such intensity that a special issue of the Journal of Educational Measurement

(Winter 1978) was devoted entirely to standard setting. Many studies have been carried out as well with varying and mixed results. Gene Glass (1978) a leading critic of standard setting contended that they were a meaningless attempt to make objective a matter that is essentially judgmental and arbitrary. Proponents while agreeing that the methods are judgmental and arbitrary to some extent, argued that these should not be taken to mean capriciousness. They further argued for the use of normative data to be used by judges as supplemental information in the standard setting exercise as a way to reduce the arbitrariness and thus set more defensible standards. Most of the work in standard setting has been a priori. That is judgments were made without knowing how the students actually performed. Recently educational reseachers have advocated the use of normative data and the process of iteration in standard setting as a "reality check" (Livingston et al 1982).

The Alberta Achievement testing program cognizent of this (in its quest to provide referenced level standards for its various jurisdictions) engaged teachers in 1983 in an iterative judgmental standard setting process, in order to provide a reference level against which jurisdictions could be measured. This study reexamined the data collected in that process. The method used was a modification of the Nedelsky (1954) procedure. Seventy-one (71) grade 9 teachers were engaged as judges to determine the standard or cut-off score for the Alberta grade 9, Social Studies Achievement

test. Judges were engaged in two sessions. Each judge was required to do two tasks:

1. To rate each of the 60 multiple choice items in terms of its curricular relevance as essential, important, adequate and questionable.
2. To inspect each item and circle the number of alternatives they thought a borderline student could eliminate as incorrect.

The second rating session only included the latter part of the first. Between sessions, judges were provided with the actual item statistical analysis for 600 students as well as their individually derived recommended standard. This gave them the opportunity to reconsider or revise their initial judgments if they so wished before the start of the second session.

It was the purpose of the study to examine six questions stemming from the use and application of normative data in this iterative modefied Nedelsky procedure relative to the Alberta grade 9 Social Studies Achievement test 1983.

6.0.1 Results

The overall standard or cut-off score set for the test was(at time2) 22 out of 60 items on the multiple choice section of the Social Studies test. There was a significant difference between overall judgments from time 1 to time 2 ($t = 15.13$ $\underline{P} < .001$). The results further indicate that cut-off scores variability differed a great deal among

individual judges and over all judges on both ocassions. Time 1 rating was more variable (interquartile range=10) with a median score of 28 whereas time 2 showed less variability (interquartile range = 4) with a lower median cut-off score (22)-although there were four outliers.

A significant relationship was also found to exist between judge judgments of estimated item difficulty at time1 and the actual test item difficulty. Finally, given the supplemental data in an iterative procedures, judges were generally more inclined to increase the rating of difficulty, which made the rated difficulty less like the actual item difficulty.

6.0.2 Conclusions

Within the context of the six research questions addressed by this study, the following conclusions can be drawn.

First, judges with similar professional characteristics(content specialty , teaching levels etc) who engaged in standard setting using the same method, set standards which are extremely variable. This was not expected. Second, normative data used in an iterative judgmental standard setting exercise do appear to influence judges' opinions. The difference in cut-off scores between rating sessions yielded statistical significence.

Third the normative data which were to aid judges in their, deliberations was used by some judges more than

others, with respect to judgment and standards for borderline students. The supplemental information appeared to separate judges into three categories: high, medium and low changers.

Fourth, it appeared as well, that when judges did change, they used the data in favour of the students by seeing the items as being less like and more difficult than the statistical difficulty level indicated. In effect they lowered the cut-off score for the borderline student. Fifth, it was observed that at time1 rating, judges estimate of item difficulty was more like the actual item difficulty, than at time2. The move appeared to be in favour of acknowledging that the borderline student would have more difficulty with all items. And sixth, there appeared to be a relationship between judges classification of item relevance and their estimates of item difficuty.

6.0.3 Implications

These findings underscore the need to examine further issues concerning standard setting exercises, particularly when engaging content specialists. The results and conclusions have both positive and negative implications for many applications of the various standard setting method. The implications are positive in that there seem to be the tendency of judges to use the normative data provided during an iterative exercise to set cut-off scores that tend to favour borderline students. This is extremely important

because it points to the fact with this supplemental information judges can be expected to set more reasonable and appropiate scores for borderline students.

The implications are negative in the sense that content specialist judgments more variable than had been anticipated. It was reasonable to expect that judges with similar professional characteristics and teaching level experiences might arrive at very similar judgments. It is therefore not sufficient to have content specialists set standards. Rather, in addition to being content specialists they should also be well trained in the method of use. Additionally, due to this variability among individual judges, cut-off scores should be based on group data rather than relying on individual judgments.

6.0.4 Recommendations

The process of setting a standard on an achievement test can only be as good as the judgments that go into it. The standard will depend on whose judgments are involved and the knowledge experience and competence the judges have in and about the method to be employed. It is therefore extremely important that the selection of judges be a meticulous exercise. They must be familiar with the test takers' knowledge and skills and that they must undergo systematic training to become relatively proficient in the standard setting procedure to be used.

Judges should continue to interact with normative data as supplemental information in a standard setting exercise. Also, they may need to be instructed in using the data.

In terms of further study, additional research should be conducted using normative data in an iterative process to ascertain more empirically additional effects of this supplemental data on judges' judgments. For example, a study could be undertaken whereby all the judges would participate in a first and second rating sessions. For the second session however, half of the judges would be given the additional information while the other half simply go through the motion without this additional data. One might then compare the differences between and within groups to see the interaction and influence of the additional normative data on judges' judgments. This method might establish more conclusively whether the judges' judgements were changed do to the exposure of additional information or whether the change was due mainly to the fact that the process was iterative.

It is also recommended that where practicable and economically feasible, judges should strive for consensus in their recommendations. Where this is found not to be practicable and feasible, then maybe some consideration should be given to using the trimmed-mean formula for individual judgments before collapsing across judges. The trimmed mean formula represents a combined score between the mean and the median. The important feature of this method is

that it gets rid of the wide dispersion or variability in judges' judgments. It does so by eliminating the highest and lowest scores from the group. The remaining scores are then averaged in the usual way to arrive at a cut-off score.

Finally a further study could be conducted whereby if the exercise is geared to setting passing scores for borderline students, the normative data should not be the test item statistics for all students. Instead, the normative data should be the test statistics of the performance of the students in the lower quartile.

REFERENCES

Andrew, B. J. and Hecht, J. T. A preliminary investigation of two procedures for setting examination standards. Educational and Psychological Measurement, 1976, 36, 45-50.

Angoff, W. Scales, norms and equivalent scores. In R. L. Thorndike (Eds.), Educational Measurement, Washington, D. C., Amerson Council in Education, 1971.

Berk, R. A. Criterion-referenced measurement: The state of the art. The John Hopkins University Press, 1978.

Block, J. H. Standards and criteria: A reponse. Journal of Educational Measurement, 1978, 15(4).

Burton, N. W. Social standards. Journal of Educational Measurement, 1978, 15, 263-271.

Conway, L. E. Discussant comments: Setting performance standards based on limited research. Florida Journal of Educational Research, XVIII, 1976.

Cross, L. H. Impara, J. C.Frary, R. B. & Jaeger R.M : A Comparison of Three Methods For Establishing Minimum Standards on The National Teacher Examinations, Journal of Educational Measurement 1984, 21(2), pp.113-129.

Degruijter, D. N. M. Accounting for the uncertainty in

performance standards, Eric Documents, ED 199 280, June, 1980.

Ebel, R. L. Essentials of Educational Measurement. Englewood Cliff, N. J., Prentice-Hall, 1972.

Flaherty, J. F. Institutional use of cut-off scores for the College level examination program, Eric Documents, ED 201 648, April, 1981.

Glass, G. Standards and criteria. Journal of Educational Measurement, 1978, 15, 237-262.

Gronlund, N. E. Constructing Achievement Test (3 Ed.), Prentice Hall Inc., Englewood Cliffs, 1982.

Gross, L. J. Standards and criteria: A response to Glass' criticism of the Nedelksy technique. Journal of Educational Measurement; 19(2), Summer, 1982.

Haladyna, T. Comments: Measurement issues related to performance standards; Florida's Journal of Educational Research XVIII, 1976.

Halpin, G. Sigmon, G. and Halpin, G. Minimum Competency standards set by three divergent groups of raters using three judgmental procedures: Implications for validity, Educational & Psychological Measurement, 1983, 43, 185-196.

Hambleton, R. On the use of cut-off scores with

criterion-referenced tests in instructional settings. Journal of Educational Measurement, 1978, 15, 277-290.

Hambleton, R. K. Test score validity and standard setting methods in criterion-referenced measurement: The state of the art; Edited by Berk, 1980.

Hambleton, R. K. et al. Issues and methods for standard-setting. ERIC Documents ED 206696, June, 1983.

Hambleton R. and Eignor, D. Competency test development, validation and standard setting. Chapter 9 in Jaeger and Tittle, 1980.

Harasym, P. H. A comparison of the Nedelsky and modified Angoff standard-setting procedure on evaluation outcome, Educational & Psychological Measurement, 1981, 41, 725-733.

Jaeger, R. M. An iterative structured judgment process for establishing standards competency tests: Theory and application. Educational Evaluation and Policy Analysis, 1982, 4, 461-475.

Jaeger, R. M. Measurement consequences of selected standard-setting Florida Journal of Educational Research, 1976, 18, 22-25.

Jaeger, R. M. and Tittle, C. K. Minimum Competency Achievement Testing, Berkeley: McCutchan Publishing, 1980.

Koffler, S. L. A. A comparison of approaches for selecting proficiency standards. Journal of Educational Measurement , 17(3), Fall, 1980.

Levin, H. M. Educational performance standards: Image or substance? Journal of Educational Measurement, 15(4) , Winter, 1978.

Linn, R. L. Demands, cautions, and suggestions for setting standards. Journal of Educational Measurement 15(4), Winter, 1982.

Livingston, S. A. and Kastrinos, W. A study of the reliability of Nedelsky's method for choosing a passing score. 1987, Research Report RR82-6, ETS, Princeton, New Jersey.

Livingston, S. A. & Zieky, M. J. Passing scores: A manual for Setting Standards of Performance on Educational & Occupational Tests, ETS, 1982.

Maguire, T. O. Standards and Guidelines in the Assessment of Student Achievement. 1983, Unpublished paper.

McLarty, J. R. Choosing passing scores for tests required for high school graduation. Eric Documents, ED 204 402, April, 1981.

Meskauskas, J. Evaluation models for criterion-referenced testing. Review of Educational Research, 1980, 6, 35-72.

Mills, C. N. A comparison of three methods of establishing cut-off scores on criterion-referenced tests, Journal of Educational Measurement, Fall 1983, 23(3), 283-292.

Mitchell, V. P. J. and Smith, R. Determining cut-off scores using Rasch Ability estimates, Eric Documents, ED 193 235, March, 1980.

Nitko, A. Criterion referencing schemes. New Directions for Testing and Measurement, 1980, 6, 35-72.

Poggio, J. P. Glasnapp, D. R. & Eros, D. S : An empirical investigation of the Angoff, Ebel, Nedelsky and contrasting group setting standard methods, Eric Documents, Ed 205 552, April, 1981.

Popham, J. W. As always provocative. Journal of Educational Measurement, 15(4), 1978.

Popham, J. W. Modern educational measurement. Prentice-Hall Inc., Englewood Cliffs, New Jersey, 1981.

Rock, D., Werts, E. and Werts, C. An empirical comparison of judgmental approaches to standard setting procedures. Research Report RR80-7, ETS, Princeton, New Jersey, 1980.

Saunders, J. C., Ryan, J. P. and Huynh, H. A comparison of two approaches to setting passing scores based on the Nedelsky procedure . Applied Psychological Measurement, 1981, 5, 209-217.

Scriven, M. How to anchor standards. Journal of Educational Measurement, 1978, 15, 273-275.

Shepard, L. A. Setting standards and living with them. Florida Journal of Educational Research, 1976, 18, 28-32.

Shepard, L. A. Technical issues in minimum competency testing. In D. C. Berliner (Ed.), Review of Research in Education, Vol. 8, Itasca, Illinois: Peacock, 1980.

Shepard, L. A. Standard setting issues and methods. Applied Psychological Measurment, 1981, 4(4), 447-467.

Skakun, E. N. and Kling, S. Comparability of methods for setting standards. Journal of Educational Measurement, 1980, 17, 229-235.

Stoker, H. W. Performance standards in competency-based education. Florida Journal of Educational Research, Vol. XVIII, 1976.

Van der, Linden W. J. Criterion-referenced measurement: Its main applications, problems, and findings. Evaluation in Education, 97-118, 1978.

Veldhuijzen, N. H. Setting cutting scores: A minimum information approach. Evaluation in Education, 1982, 5, 141-148.

Appendices

329G Michener Park
Edmonton, Alberta
T6G 1M5
(Res) 435-3197
(Bus) 432-5893

Dr. L. E. Symyrozum
Director
Student Evaluation Branch
Alberta Education
Edmonton, Alberta

Dear Dr. Symyrozum:
First allow me to introduce myself. I am an Acting Education Officer with the Bahamas' Ministry of Education currently on study leave on a Canadian Commonwealth Scholarship. At present I am an M.Ed. candidate at the University of Alberta.

For my M.Ed. thesis, I wish to examine the notion of standard setting in the Alberta context, using, a modification of the Nedelsky (1954) and Ebel (1972) methods respectively. Consequently, I need the use of some normative data for my study (data specific to 1982 Grade 3 math achievement test).

As a result of my discussion with Dr. T. O. Maguire, I was advised that the following information will be of use to me in this study:

- item data (frequencies and percentages of responses to alternatives)
- score distribution
- item difficulty
- copy of exam papers
- test blue print
- curriculum specification
- reports to schools and jurisdictions (identification of schools or jurisdictions is not necessary).

I shall be grateful therefore, if you will be kind enough to let me have access to and use of this information.

Please accept my thanks in anticipation of your generosity and please note that a copy of my proposal is enclosed for your information.

Sincerely yours

LeRoy Sumner

enc

APPENDIX A

Item Rating Form Instructions

The major purpose of this exercise is to assist us in indentifying a test score which would indicate a minimum acceptable level of achievement, assuming adequate delivery of the program. Another purpose is to collect information on the importance of each item to the program as teachers perceive it.

Step 1: Please print your name and marker number in the spaces provided.

Step 2: Examine each item in the multiple-choice section of the test. Determine how important the knowledge or skill which that item tests is in your Social Studies program. Indicate your judgments by circling the appropriate letter under importance, next to the item number.

Step 3: Please imagine a student at the 15th percentile in terms of ability. this student would be third from the bottom of the class in ability. Then decide how many of the alternatives for that item that the typical 15th percentile student would eliminate as obviously wrong. Indicate your judgment by circling the appropriate number under First Judgment next to the item number.

Repeat steps 2 and 3 for all 60 items.

The Revised Judgment portion of the sheet is reserved for possible later use.

Item Rating Form

Name:_____Number:____

Item	Importance*	No. of Alternatives Eliminated by 15th %ile Student: First Judgment	Revised Judgment
1	E I A Q	0 1 2 3	0 1 2 3
2	E I A Q	0 1 2 3	0 1 2 3
3	E I A Q	0 1 2 3	0 1 2 3
4	E I A Q	0 1 2 3	0 1 2 3
5	E I A Q	0 1 2 3	0 1 2 3
6	E I A Q	0 1 2 3	0 1 2 3
7	E I A Q	0 1 2 3	0 1 2 3
8	E I A Q	0 1 2 3	0 1 2 3
9	E I A Q	0 1 2 3	0 1 2 3
10	E I A Q	0 1 2 3	0 1 2 3
11	E I A Q	0 1 2 3	0 1 2 3
12	E I A Q	0 1 2 3	0 1 2 3
13	E I A Q	0 1 2 3	0 1 2 3
14	E I A Q	0 1 2 3	0 1 2 3
15	E I A Q	0 1 2 3	0 1 2 3
16	E I A Q	0 1 2 3	0 1 2 3
17	E I A Q	0 1 2 3	0 1 2 3
18	E I A Q	0 1 2 3	0 1 2 3
19	E I A Q	0 1 2 3	0 1 2 3
20	E I A Q	0 1 2 3	0 1 2 3
21	E I A Q	0 1 2 3	0 1 2 3
22	E I A Q	0 1 2 3	0 1 2 3

*E=Essential I=Important A=Acceptable Q=Questionable

Item	Importance*	No. of Alternatives Eliminated by 15th %ile Student: First Judgment	Revised Judgment
23	E I A Q	0 1 2 3	0 1 2 3
24	E I A Q	0 1 2 3	0 1 2 3
25	E I A Q	0 1 2 3	0 1 2 3
26	E I A Q	0 1 2 3	0 1 2 3
27	E I A Q	0 1 2 3	0 1 2 3
28	E I A Q	0 1 2 3	0 1 2 3
29	E I A Q	0 1 2 3	0 1 2 3
30	E I A Q	0 1 2 3	0 1 2 3
31	E I A Q	0 1 2 3	0 1 2 3
32	E I A Q	0 1 2 3	0 1 2 3
33	E I A Q	0 1 2 3	0 1 2 3
34	E I A Q	0 1 2 3	0 1 2 3
35	E I A Q	0 1 2 3	0 1 2 3
36	E I A Q	0 1 2 3	0 1 2 3
37	E I A Q	0 1 2 3	0 1 2 3
38	E I A Q	0 1 2 3	0 1 2 3
39	E I A Q	0 1 2 3	0 1 2 3
40	E I A Q	0 1 2 3	0 1 2 3
41	E I A Q	0 1 2 3	0 1 2 3
42	E I A Q	0 1 2 3	0 1 2 3
43	E I A Q	0 1 2 3	0 1 2 3
44	E I A Q	0 1 2 3	0 1 2 3
45	E I A Q	0 1 2 3	0 1 2 3
46	E I A Q	0 1 2 3	0 1 2 3

*E=Essential I=Important A=Acceptable Q=Questionable

Item	Importance*	No. of Alternatives Eliminated by 15th %ile Student First Judgment	Revised Judgment
47	E I A Q	0 1 2 3	0 1 2 3
48	E I A Q	0 1 2 3	0 1 2 3
49	E I A Q	0 1 2 3	0 1 2 3
50	E I A Q	0 1 2 3	0 1 2 3
51	E I A Q	0 1 2 3	0 1 2 3
52	E I A Q	0 1 2 3	0 1 2 3
53	E I A Q	0 1 2 3	0 1 2 3
54	E I A Q	0 1 2 3	0 1 2 3
55	E I A Q	0 1 2 3	0 1 2 3
56	E I A Q	0 1 2 3	0 1 2 3
57	E I A Q	0 1 2 3	0 1 2 3
58	E I A Q	0 1 2 3	0 1 2 3
59	E I A Q	0 1 2 3	0 1 2 3
60	E I A Q	0 1 2 3	0 1 2 3

*E=Essential I=Important A=Acceptable Q=Questionable

APPENDIX B

Observed Item Response Frequencies

Item #	Key	Percentage of All Students Choosing Alternative			
		A	B	C	D
1	C	9	10	77	4
2	B	37	34	17	12
3	C	8	8	41	43
4	D	7	5	6	83
5	D	17	6	10	66
6	C	3	11	66	20
7	A	61	8	18	13
8	D	7	12	5	76
9	C	23	4	58	15
10	B	5	87	4	4
11	B	8	74	11	7
12	B	9	46	27	18
13	D	11	3	5	81
14	C	12	18	46	23
15	C	5	10	70	16
16	A	54	11	24	10
17	C	22	17	60	1
18	D	16	24	9	51
19	B	19	62	13	7
20	D	3	4	21	71
21	D	11	11	4	74
22	A	79	5	6	11
23	C	2	5	85	8
24	C	40	13	43	4
25	A	60	23	9	9
26	D	18	11	14	57
27	A	62	27	6	5
28	B	3	78	14	5
29	A	52	7	15	27
30	C	19	13	43	25
31	B	14	50	13	23
32	B	10	62	12	16
33	D	4	11	18	66
34	D	18	7	35	41
35	C	5	8	76	10
36	B	9	69	16	6
37	B	27	43	12	17
38	A	75	6	14	5
39	B	2	3	9	36
40	C	16	18	48	18
41	D	10	5	34	51
42	D	7	9	13	72

Table 5 (Continued)

Item #	Key	Percentage of All Students Choosing Alternative			
		A	B	C	D
43	D	8	18	9	65
44	A	66	9	13	12
45	D	21	13	17	49
46	B	21	50	20	9
47	A	75	10	10	5
48	A	73	11	10	6
49	A	34	25	25	16
50	B	7	66	11	17
51	A	51	25	12	12
52	C	9	18	65	8
53	A	49	13	18	19
54	B	9	54	31	6
55	D	9	6	16	69
56	C	8	9	74	8
57	C	11	11	58	20
58	D	9	12	17	62
59	B	20	59	11	11
60	A	41	15	15	29

Appendix C

Low Changes

Judge	Change Score	Item Changed	f	Proportion in each category			
				E	I	A	Q
28	57	3		0	2	1	0
				.000	.67	.33	.000
38	57	3	3	1	2	0	0
				.33	.67	.000	.000
63	57	3		1	2	0	0
				.33	.67	.000	.000
43	56	4	1	0	1	3	0
				.000	.25	.75	.000
39	55	5	1	2	2	0	1
				.40	.40	.000	.10
41	53	7	1	2	4	1	0
				.285	.571	.142	.000
22	52	8	1	0	7	1	0
				.000	.875	.125	.000
47	51	9	1	2	1	6	0
				.222	.111	.667	.000
40	50	10	1	2	6	2	0
				.20	.60	.20	.000
7	48	12	1	1	11	0	0
				.083	.917	.000	.000
20	47	13		2	7	4	0
			2	.167	.583	.333	.000
50	47	13		4	5	3	1
				.307	.384	.230	.076
23	45	15		7	8	0	0
				.467	.533	.000	.000
27	45	15	3	4	9	2	0
				.267	.60	.133	.000
57	45	15		11	4	0	0
				.733	.267	.000	.000

Appendix C

Low Changes

19	43	17		6	9		0
				.352	.529	.117	.000
54	43	17	3	3	11	3	0
				.176	.647	.176	.000
60	43	17		5	7	4	1
				.294	.411	.235	.058
35	42	18			6	6	1
			2	.278	.333	.333	.056
51	42	18		10	7	1	0
				.556	.339	.056	.000
N=20							
				68	111	39	4

Appendix D

Medium Changes

Judges	Change Score	Items Changed	f	Proportion in each Category			
				E	I	A	Q
3	41	19	1	6	11	2	0
				.315	.578	.105	.000
20	39	21	1	3	14	4	0
				.143	.667	.190	.000
46	37	23	1	6	11	5	1
				.260	.478	.217	.043
21	36	24	1	4	12	8	0
				.167	.500	.333	.000
29	35	25		20	2	3	0
			2	.800	.080	.120	.000
34	35	25		9	7	9	0
				.360	.280	.360	.000
32	34	26	1	7	14	4	1
				.269	.538	.154	.038
26	33	27		15	12	0	0
			4	.556	.444	.000	.000
49	33	27		13	13	1	0
				.481	.481	.037	.000
33	32	28		7	16	5	0
			2	.250	.571	.179	.000
65	32	28		4	11	11	2
				4	11	11	2
				.143	.393	.393	.071
15	31	29		6	10	13	0
			2	.207	.345	.448	.000
68	31	29		9	16	3	1
				.310	.552	.103	.034
11	30	30	1	4	24	2	0
				.133	.800	.067	.000
8	29	31		12	18	1	0
				.387	.581	.032	.000
42	29	31		6	14	11	0
			5	.194	.452	.354	.000
51	29	31		5	16	10	0
				.161	.516	.323	.000
69	29	31		15	15	2	0
				.452	.484	.064	.000
62	28	32		9	17	6	0
			2	.281	.531	.188	.000
70	28	32		20	9	3	0
				.625	.281	.093	.000

Judges	Change Score	Items Changed	f	Proportion in each Category			
				E	I	A	Q
2	27	33		19	9	4	1
				.575	.272	121	.030
14	27	33		7	16	8	2
				.212	.485	.242	.061
44	27	33		12	19	2	0
			4	.362	.576	.061	.000
66	27	33		3	28	1	1
				.091	.848	.030	.030
53	26	34	1	6	16	11	1
				.176	.471	321	.029
1	25	35			18	11	1
			2	.171	.514	.314	.028
45	25	35		10	25	4	0
				.285	.714	.114	.000
18	24	36		5	30	1	0
				.138	.833	.027	.000
24	24	36	3	32	3	1	0
				.889	.083	.028	.000
59	24	36		2	29	5	0
				.056	.806	.139	.000

N=30

Appendix E

High Changes

Judge	Change Score	Item Changed	f	Proportion in each category			
				E	I	A	Q
10	23	37		21	13	3	0
				.567	.352	.081	.000
16	23	37	3	1	16	15	5
				.027	.432	.405	.135
17	23	37		17	18	2	0
				.460	.486	.054	.000
5	22	38		0	8	26	4
				.000	.210	.684	.105
13	22	38	4	11	19	8	0
				.289	.50	.211	.000
64	22	38		0	27	10	1
				.000	.711	.263	.026
67	22	38		13	22	3	0
				.342	.579	.079	.000
36	21	39		5	19	11	4
				.128	.487	.282	.103
48	21	39	3	7	26	6	0
				.179	.667	.154	.000
56	21	39		26	10	3	0
				.667	.256	.077	.000
6	20	40		9	16	13	2
				.225	.40	.325	.050
			2				
37	20	40		9	20	11	0
				.225	.50	.275	.000
61	19	41	1	9	24	6	2
				.219	.585	.146	.048
4	18	42	2	23	15	4	0
				.578	.357	.095	.000
25	18	42		11	17	12	2
				.262	.405	.286	.047
12	17	43	1	17	18	7	1
				.395	.419	.162	.023
31	16	44		28	12	4	0
				.637	.273	.090	.000
71	16	44		5	24	13	2
				.113	.545	.295	.045

Judge	Change Score	Item Changed	f	Proportion in each category			
				E	I	A	Q
			2				
9	15	45	1	6	11	20	
				.133	.244	.444	.178
55	14	46	1	2	26	11	7
				.043	.565	.239	.152
58	11	49	1	20	20	7	2
				.408	.408	.143	.040
N=21				240	381	195	40